UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Implementação de uma metodologia para análise metagenómica do microbioma humano em amostras com baixa biomassa de microrganismos

Dina Isabel Filipe Carpinteiro Durão

Mestrado em Biologia Humana e Ambiente

Dissertação orientada por:
Doutor Luís Vieira
Profª Doutora Deodália Dias

2020

Agradecimentos

Agradeço ao Doutor Luís Vieira pelo incentivo a fazer mais e melhor sempre. Tenho a sorte de ter um chefe e um amigo na mesma pessoa.

Agradeço aos colegas de trabalho, pois muitas vezes tiveram de avançar com a rotina do laboratório para que eu pudesse dedicar-me a este projeto.

Agradeço aos meus pais que sempre fizeram o melhor por mim e foram muitas vezes o meu pilar de ajuda com os meus filhos.

Agradeço aos meus filhos Pedro e Carol, e ao meu marido Carlos Durão, pelo suporte emocional que tantas vezes precisei.

Obrigada a todos

As referências bibliográficas foram citadas de acordo com o estilo de citações da revista *Nature*.

Resumo

O corpo humano é colonizado por um grande conjunto de microrganismos que vivem em simbiose com o homem. Desta colonização fazem parte bactérias, fungos, vírus e protozoários, que se espalham ao longo das superfícies externas e internas do corpo, criando microbiomas com características e funções específicas. Os estudos do microbioma vieram beneficiar de forma significativa dos avanços das novas tecnologias de sequenciação ("next-generation sequencing"), as quais permitem sequenciar elevadas quantidades de DNA a um custo muito reduzido. No entanto, este tipo de estudos também acarreta algumas dificuldades, nomeadamente a presença de uma pequena biomassa de microrganismos em alguns locais do corpo e a contaminação com estirpes ambientais durante a colheita e processamento das amostras. Neste trabalho foi implementada uma metodologia para analise de contaminantes em amostras com baixa biomassa. Para este objectivo, foi extraído DNA genómico de 34 amostras de tecido renal tumoral em 9 séries de extração, tendo sido incluído 1 controlo branco em cada procedimento de extracção. Os DNAs foram amplificados por *nested* PCR para a região V3 e V4 do gene 16S rRNA, tendo sido incluído um controlo negativo da PCR por cada conjunto de amostras amplificadas. Os amplicões do gene 16S rRNA compreenderam um total de 96 amostras e controlos, incluindo duplicados que foram sequenciados em plataforma Illumina, tendo-se obtido um total de 47.3 milhões de reads. Estas reads foram tratadas e processadas usando a plataforma bioinformática QIIME2. Após a filtragem e o denoising das reads, a análise taxonómica revelou a presença de 19 filos, 27 classes, 55ordens, 75 famílias, 89 géneros e 114 espécies distintas em amostras e controlos. Usando o pacote Decontam do R, foram identificados 35 géneros contaminantes nos controlos da extracção e 17 géneros contaminantes foram identificados nos controlos da PCR. A proporção de contaminantes nas amostras de DNA variou entre 0,01% e 24,8%. Entre estes, foram detectados 9 géneros que não estão normalmente presentes como contaminantes em estudos de metagenómica, pelo que a sua detecção neste estudo permitirá analisar com cautela os resultados de futuras amostras com estes géneros. Uma vez que a maioria dos géneros contaminantes foi detectada nos controlos da extracção, deve ser dada especial atenção à pureza e manipulação dos reagentes utilizados na preparação das amostras. Dos géneros contaminantes detectados, 9 não foram referidos em outros estudos como contaminantes, indicando que a presença destes microorganismos em estudos de microbioma futuros, devem ser interpretados com precaução. Concluímos que as estirpes contaminantes são um problema sério em estudos de microbioma, em amostras com baixa biomassa e que a metodologia apresentada aqui é uma abordagem eficiente para detectar e quantificar essas estirpes contaminantes.

Palavras-chave: Sequenciação de próxima-geração, bioinformática, metagenómica, gene *16S rRNA*, contaminantes

Abstract

The human body is colonized by a large group of microorganisms that live in symbiosis with humans. This colonization includes bacteria, fungi, viruses and protozoa, which spread along the external and internal surfaces of the body, creating microbiomes with specific characteristics and functions. Studies of microbiome benefited significantly from advances in new sequencing technologies ("next generation sequencing"), as these allow sequencing high amounts of DNA at low costs. However, this type of studies also brings difficulties, namely the presence of small biomasses of microorganisms in some parts of the body, which poses technical limits to their detection, and contamination with environmental strains during the collection and processing of samples, which may interfere with the true taxonomic profile of microorganism communities. In this work, a methodology for the analysis of contaminants in low biomass samples was implemented. For this purpose, nine DNA extraction procedures were performed, comprising a total of 34 tumoral renal tissue samples, each of which included a reagent only (negative) control. The DNAs were amplified by nested PCR for a V3-V4 region of the 16S rRNA gene, and a PCR negative control was included in each reaction set. The 16S rRNA amplicons comprising a total of duplicate 96 samples and controls, including duplicates, were sequenced on an Illumina platform, producing a total of 47.3 million reads. These reads were treated and processed using QIIME2 microbiome bioinformatics platform. After filtering and denoising of reads, taxonomy analysis revealed the presence of 19 phyla, 27 classes, 55 orders, 75 families, 89 genera and 114 distinct species in samples and controls. Using the Decontam R package, 35 contaminating genera were identified in the DNA extraction controls and 17 contaminating genera were found in the PCR controls. The proportion of contaminants in DNA samples varied between 0.01% and 24,8%. Since the majority of the contaminating genera were detected in the extraction controls, particular attention should be paid to the preparation and manipulation of the reagents used in DNA extraction procedures. Among the contaminating genera, 9 were not present as contaminants in other studies, indicating that the presence of these microorganisms in future microbiome studies should be interpreted with caution. We conclude that contaminating strains are a serious problem in low biomass microbiome studies and that the methodology presented here is an efficient approach to detect and quantify those strains.

Keywords: Next-generation sequencing, bioinformatics, metagenomics, 16S rRNA gene, contaminants

Índice

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ASV – Amplicon Sequence Variant

BR- Broad Range

HT1 – Hybridization Buffer

NGS – Next Generation Sequencing

Pb – Pares de base

PCA – Principal Component Analysis

PCR – Polimerase Chain Reaction

QIIME2 – Quantitative Insights Into Microbial Ecology

SDS – Dodecil sulfato de sódio

STE – Solução de Sodium Chloride-TRIS-EDTA buffer

TE - Solução de Tris-EDTA

1 Introdução

1.1 O microbioma humano e o seu papel na saúde e na doença

À luz do conhecimento atual, é possível afirmar com alguma segurança que todos os tipos de biomas que podem ser encontrados no planeta Terra, possuem uma comunidade própria de microrganismos com funções específicas. Estes biomas podem ser tão diversos quanto as águas profundas dos oceanos, as raízes das plantas ou a cavidade intestinal do organismo humano. Neste último exemplo, o conjunto de microrganismos como bactérias, vírus, fungos e protozoários, que mantêm uma relação de simbiose com o homem, é designado por **microbioma humano.**

O termo microbioma foi definido pela primeira vez por Joshua Lederberg, sendo muitas vezes referido como um "segundo" genoma humano [1]. Estima-se que o microbioma humano seja composto por cerca de 10 a 100 triliões de microrganismos, dispersos pelas superfícies externa e interna do corpo humano, e que o número de células humanas constitui menos de metade do número total de células do organismo. Assim, estas comunidades de microrganismos ocupam os mais variados locais anatómicos do corpo humano (**figura 1**), criando micro-habitats e desempenhando funções específicas de acordo com o órgão ou tecido que colonizam [2]. Devido à relação simbiótica que estabelecem com as células humanas, os microrganismos que compõem o microbioma não têm à partida características patogénicas que possam ser causadores diretas de doença. No entanto, a variação qualitativa ou quantitativa das estirpes que participam na composição do microbioma humano, podem estar associadas ao desenvolvimento de certas doenças ou afetar a eficácia de determinados agentes terapêuticos. Neste contexto, a caracterização da composição taxonómica do microbioma e do seu perfil funcional tem-se revelado de grande importância na saúde humana, uma vez que vários estudos têm indicado associações entre o microbioma e a manifestação de certas doenças [3]. Em circunstâncias normais, estas comunidades bacterianas vivem em equilíbrio com o nosso organismo. No entanto, diversos fatores externos, como o uso de antibióticos, hábitos alimentares ou a exposição a fatores ambientais, podem conduzir a um desequilíbrio chamado **disbiose**. A disbiose pode originar estados de doença como obesidade, diabetes, alterações da flora vaginal, alterações bacterianas da pele, entre outros [2].

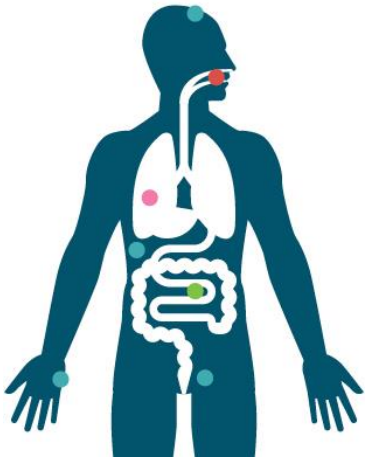


Figura 1. Representação de micro-habitats de comunidades bacterianas presentes em diversos locais do corpo humano. Os locais externos estão indicados com círculos azuis e os locais internos estão representados com círculos de outras cores. Adaptado de https://www.the-scientist.com/features/the-mycobiome-34129

1.2 Origem e desenvolvimento do microbioma humano

Atualmente sabe-se que o microbioma humano é adquirido quase na sua totalidade aquando do nascimento, havendo transmissão de microrganismos através da progenitora. No caso do parto normal, o recém-nascido recebe o microbioma da cavidade vaginal enquanto nas cesarianas é transmitido o microbioma da pele [2]. A relevância desta diferença no estabelecimento do microbioma de cada indivíduo e a sua influência no contexto da saúde e da doença, é atualmente uma área em discussão na comunidade científica. Alguns autores defendem que o microbioma também pode ser adquirido durante a gestação do feto. O meio "ambiental" que envolve o feto durante a gestação, ou seja, a placenta e o líquido amniótico, foram considerados durante muito tempo como um ambiente gestacional estéril, mas alguns estudos recentes sugerem que a aquisição do microbioma acontece durante o desenvolvimento gestacional, passando parte do microbioma da mãe para o filho através do cordão umbilical [4] . Contudo, esta hipótese não foi suportada em alguns estudos, nos quais não foi possível identificar com segurança a presença de microrganismos no líquido amniótico [5]. O microbioma que irá fazer parte de toda a vida adulta só estabiliza aos cerca de 3 anos de idade. Isto deve-se a alterações ao longo dos primeiros anos de vida, nomeadamente alterações da composição da microbiota intestinal, de acordo com a amamentação materna e a dieta alimentar [6].

1.3 Tecnologias de estudo do microbioma humano

Nos últimos anos, o avanço das novas tecnologias de sequenciação (sequenciação de próxima geração ou "next-generation sequencing", NGS), tem permitido demonstrar a importância do papel dos microrganismos e a sua interação com a saúde/doença humana. Estes avanços levaram à criação de um projeto intitulado "Human Microbiome Project", cujo objetivo foi caracterizar a diversidade microbiota de indivíduos humanos saudáveis, presente em várias partes do corpo, e assim conhecer melhor a interação do microbioma com o hospedeiro [7]. Com este grande projeto, as tecnologias de NGS vieram dar um grande impulso aos estudos do microbioma, uma vez que permitiram sequenciar uma quantidade muito elevada de fragmentos de DNA a um custo cada vez mais reduzido. Foi assim possível caracterizar comunidades de microrganismos com elevada profundidade e assim detetar estirpes minoritárias, ou novas espécies, que exercem um papel funcional importante nessas comunidades.

Existem várias plataformas de NGS com grande capacidade de sequenciação massiva. A plataforma de sequenciação da Illumina desenvolvida em 2006 segue o princípio da química de sequenciação por síntese, com recurso a DNA polimerase e a nucleótidos marcados com fluoróforos diferentes. Os fragmentos de DNA são ligados a sequências adaptadoras nas suas extremidades 5´, desnaturados em DNA de cadeia simples e hibridados a oligos complementares na *flowcell*, onde ocorre a sequenciação, deixando a extremidade 3´ disponível para síntese da cadeia. Segue-se a amplificação dos mesmos, através de uma estrutura em forma de ponte ("*bridge amplification*"), que gera *clusters* com fragmentos de DNA clonais prontos a serem sequenciados. A cada ciclo de sequenciação apenas 1 nucleótido terminador marcado com fluorescência (correspondente à base complementar ao fragmento que está a ser sequenciado) é detetado pelo sequenciador em cada *cluster*. No final de cada ciclo ocorre a clivagem do nucleótido terminador marcado com fluorescência, e é efetuada uma lavagem que remove reagentes excedentes assim como o fluoróforo do nucleótido incorporado no ciclo anterior, dando continuidade à sequenciação [8].

1.4 A metagenómica como metodologia de estudo do microbioma

A metodologia mais utilizada para caracterizar os microrganismos presentes em amostras de microbioma é a **metagenómica**. O termo "*metagenomics*" foi descrito em 1998 por Jo Hanselman após ter efetuado estudos de genomas em amostras de microflora de solos [9]. Esta metodologia permite a análise de fragmentos de DNA obtidos diretamente da extração das amostras ambientais, sem que seja necessário fazer culturas laboratoriais, o que veio facilitar grandemente aquele tipo de estudos quando se sabe que a maioria dos microrganismos não cresce em ambiente laboratorial [10].

Atualmente, a principal ferramenta de abordagem metagenómica utilizada para a caracterização do microbioma humano é a sequenciação do gene marcador 16S rRNA, que permite identificar as diferentes espécies presentes nas amostras [11]. Este gene tem sido amplamente utilizado por permitir analisar relações filogenéticas entre taxa distantes [12]. O gene 16S rRNA codifica para a componente menor do ribossoma, tem aproximadamente 1500 pares de bases (pb) de comprimento e compreende 9 regiões variáveis (V1 a V9) intercaladas com 9 regiões altamente conservadas ao longo da sequência, conforme representado pela **figura 2**. É considerado um gene conservado evolutivamente, e muito utilizado em análises filogenéticas de procariotas. Assim, a sequência do gene que codifica o 16S rRNA pode ser usada para identificar diferentes espécies dentro de uma comunidade bacteriana, recorrendo a *primers* universais que se ligam às regiões conservadas e amplificam as regiões variáveis.

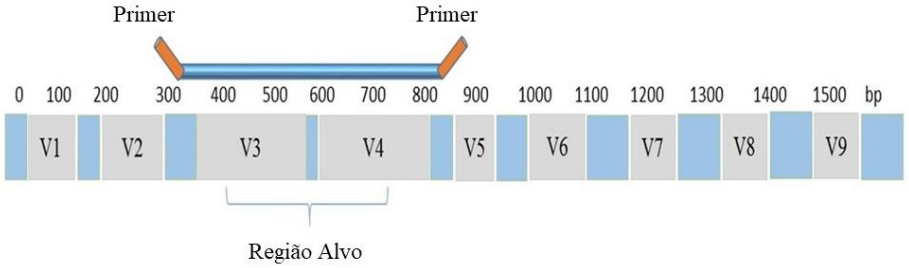


Figura 2. Representação esquemática do gene 16S rRNA, com as regiões conservadas (a azul) e as regiões variáveis (a cinzento), exemplificando a amplificação de um segmento compreendendo as regiões variáveis V3 e V4. Adaptado de https://www.lcsciences.com/discovery/applications/genomics/16s-mobile

1.5 Análise bioinformática de sequências do gene 16S rRNA

Os dados de metagenómica gerados após a sequenciação do gene 16S rRNA, contêm uma grande quantidade de informação que necessita de ferramentas adequadas para ser analisada. As sequências obtidas pelo equipamento de sequenciação (por exemplo ficheiros com a extensão *.fastq)* são numa primeira fase submetidas a uma análise primária para controlo de qualidade das *reads* obtidas. Habitualmente, este controlo é efetuado no próprio *software* do equipamento de sequenciação no caso do equipamento *MiSeq, os softwares MiSeq Control Software e o Sequencing Analysis Viewer,* permitindo a visualização geral das métricas de qualidade obtidas na corrida de sequenciação.

Seguindo para as análises secundárias, existem várias *pipelines* descritas na literatura para análise de dados metagenómicos No entanto, o QIIME2 é um dos *softwares* gratuitos mais utilizados, disponível *online* no *site* https://qiime2.org/ e que funciona em ambiente miniconda [13] . A análise de dados do gene 16S rRNA inicia-se com a remoção de sequências com baixa qualidade, a junção de pares de sequências numa única sequência, a remoção de quimeras, e a atribuição dos *amplicon sequence variants* (ASVs) a cada amostra [14]. A partir desta "limpeza" seguem-se as análises de diversidade filogenética e de caracterização taxonómica. A identificação de estirpes ambientais contaminantes pode ser efectuada com o Decontam, um pacote que funciona em R [15], e que permite a identificação de sequências contaminantes através da prevalência (presença/ausência) de cada ASV nas amostras em estudo, comparando com a prevalência dos mesmos em controlos negativos.

1.6 As contaminações ambientais nos estudos de metagenómica

Uma das dificuldades existentes nos estudos do microbioma humano é o facto de as amostras recolhidas de determinadas partes anatómicas, conterem uma baixa biomassa microbiana, tornando-se num grande desafio a caracterização exacta das comunidades microbianas. Esta evidência sugere a necessidade de se adotar uma estratégia com base na metodologia de *Polymerase Chain Reaction* (PCR), que seja capaz de detetar sequências de origem microbiana com uma maior sensibilidade. No entanto, uma estratégia mais sensível de amplificação pode conduzir à seleção de sequências provenientes de microrganismos do ambiente circundante (por exemplo, reagentes, consumíveis, superfícies de trabalho, ventilações de ar, equipamento de proteção individual, etc.) e do próprio operador, que importa detectar e conhecer do ponto de vista taxonómico.

As contaminações ambientais podem introduzir resultados falsos positivos, o que pode levar a uma caracterização incompleta e/ou distorcida das amostras em estudo, nomeadamente a nível da distribuição taxonómica e respectiva frequência. As contaminações podem acontecer nas diferentes fases do processo, principalmente ao nível da colheita e armazenamento de amostras, extração de DNA ou reação da PCR. Alguns estudos têm relatado a presença de DNA contaminante, de origem microbiana, em água usada para técnicas de biologia molecular, kits de extração de DNA e reagentes da PCR [16]. É assim relevante implementar uma metodologia de trabalho laboratorial e respectiva análise de dados para conhecer e caracterizar o impacto destas contaminações em estudos de microbioma humano, que envolvam amostras com baixa biomassa de microrganismos. Esta metodologia será importante em estudos aplicados a fetos (como referido acima) ou em outros estudos que envolvam locais anatómicos em que a presença de microorganismos é nula ou residual, como o caso de órgãos internos, como o rim humano.

2 Objetivos

2.1 Objetivo Geral

Este trabalho tem como objetivo geral implementar uma metodologia de análise metagenómica que permita avaliar e quantificar a influência da contaminação exógena no estudo de amostras humanas com baixa biomassa de microrganismos.

2.2 Objectivos Específicos

Os objetivos específicos são os seguintes:

Componente "wet lab": corresponde à fase laboratorial do protocolo para amplificação e sequenciação do gene 16S rRNA, com condições que permitam avaliar o risco de contaminações ambientais por microrganismos, nas amostras e em reagentes. Resumidamente esta componente terá as seguintes etapas:

- Extração de DNA e avaliação qualitativa/quantitativa, com introdução de controlos brancos
- Amplificação do gene 16S rRNA e de controlos negativos
- Preparação das bibliotecas de DNA
- Sequenciação paralela massiva (MiSeq, Illumina)

Componente "dry lab": corresponde à análise bioinformática, com o objetivo de estabelecer uma caracterização taxonómica do microbioma humano, com identificação e quantificação de estirpes ambientais contaminantes. Esta componente compreenderá as seguintes fases:

- Análise de qualidade e tratamento das sequências
- Análises de diversidade
- Identificação e caracterização taxonómica
- Caracterização taxonómica e quantificação de estirpes contaminantes

3 Materiais e Métodos

3.1 Amostras

Para o estudo de amostras de baixa biomassa de microrganismos, foram selecionadas 34 amostras de tecido renal tumoral humano, pertencentes ao biobanco do Departamento de Genética Humana do INSA, e que foram devidamente anonimizadas. Estas amostras eram provenientes de nefrectomias de doentes pediátricos oncológicos, e apresentavam-se em fragmentos de tecido renal com dimensões variáveis entre 2 a 3cm (**figura 3**), armazenados em ultracongeladores a -80ºC. A proposta deste estudo foi previamente submetida à Comissão de Ética em Saúde do INSA, tendo sido obtido um parecer favorável à sua realização.

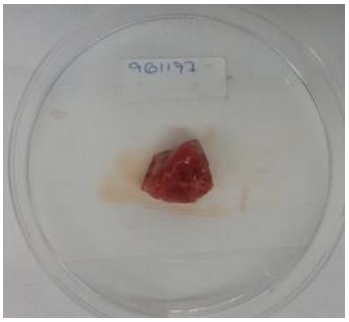


Figura 3. Exemplo de fragmento de tecido renal para extração de DNA

3.2 Extração e avaliação quantitativa/qualitativa do DNA genómico

Foram extraídos DNAs de 34 amostras de tecido renal tumoral, divididas por 9 séries de extração, onde foram incluídos controlos negativos em cada série para identificação *a posteriori* de eventuais contaminações ambientais e/ou de reagentes que possam ter sido introduzidas durante este procedimento. A extração de DNA genómico foi efetuada manualmente com recurso a soluções de proteínase K e dodecilsulfato de sódio (SDS) 20% para digestão das membranas celular/nuclear e fração proteica das células renais. O lisado celular foi purificado pelo método fenol-clorofórmio seguido de precipitação com etanol a 70% e ressuspensão em solução Tris-EDTA (TE) 1X pH7.5, de acordo com o protocolo descrito na **tabela I**.

Tabela I. Procedimento experimental de extração manual de DNA genómico a partir de fragmentos de biópsias tumorais

Etapa	Procedimento
1	Retirar, com um bisturi, um fragmento de tecido renal, cortar em frações mais pequenas, e colocar num tubo microtubo de 1,5 ml.
2	Ressuspender os fragmentos tumorais sem 250µl de STE (*Chorion Buffer*) e transferir para um microtubo de 2 ml.
3	Repetir o passo anterior adicionando mais 250µl de STE.
4	Lavar a ponta com 200µl de STE para obter no final 700µl de suspensão de células.
5	Adicionar 17,5µl de SDS 20% e 7µl de proteínase K. Vortexar a mistura para
6	homogeneizar.
7	Incubar a 55ºC, durante 3horas, num bloco térmico. Após incubação colocar o tubo em gelo.
8	Adicionar 700*ul* de fenol ao tubo e agitar por inversão.
9	Centrifugar 5min a 13000rpm.
10	Transferir a fase aquosa (sobrenadante) para outro microtubo e adicionar 700µl de fenol.
11	Centrifugar 5min a 13000rpm.
12	Transferir a fase aquosa para outro microtubo e adicionar 700µl de clorofórmio. Agitar por inversão.
13	Centrifugar 5min a 13000rpm.
14	Transferir a fase aquosa para outro microtubo e adicionar 70µl de acetato de sódio 3M pH 5.2 e 1400µl de etanol absoluto (refrigerado a -20°C).
15	Inverter o tubo várias vezes lentamente até se observar a precipitação de filamentos de DNA.
16	Centrifugar 15min a 13000rpm a 4ºC. Descartar o sobrenadante.
17	Adicionar 1ml de etanol a 70% para lavar o pellet.
18	Centrifugar 15min a 13000rpm a 4ºC. Descartar o sobrenadante.
19	Secar o tubo com o DNA no Savant durante 15min.
20	Ressuspender o DNA em 100µl de TE 1X pH 7.5.
21	Colocar o tubo a 55ºC num bloco térmico até dissolução completa do DNA.

A concentração e o grau de pureza do DNA extraído das amostras foram determinados por espectroscopia de absorção, com recurso a um espectrofotómetro *NanoVuePlus* (*GE Helthcare*). As absorvâncias foram determinadas nos comprimentos de onda de 230, 260, 280 e 320 nm e os rácios A260/A280 e A260/A230 foram usados para determinar o grau de pureza do DNA relativamente à contaminação com proteínas e fenol, respetivamente. Os DNAs foram também quantificados usando o fluorímetro *Qubit®3.0* (*Thermo Fisher Scientific*), e o kit de quantificação *Broad Range (BR*), por forma a obter os valores de concentração do DNA em cadeia dupla. Os controlos de extração não foram sujeitos a quantificação de DNA. Os DNAs foram diluídos para uma concentração final de 100 ng/µl em tampão TE 1X pH 7.5 e armazenados a 4ºC até à sua utilização posterior.

A avaliação da integridade/qualidade do DNA foi realizada através de eletroforese em gel de agarose. O gel de agarose foi preparado a uma concentração de 2% e incorporado com brometo de etídeo. As amostras, em conjunto com o marcador de peso molecular de 100 pb (Bioron), correram numa tina de eletroforese com tampão Tris-Borato-EDTA 1X, a uma voltagem de 80 volts (V), durante aproximadamente 1hora. O DNA foi visualizado sob luz ultravioleta em transiluminador, e as imagens foram obtidas utilizando o equipamento *FireReader* (*Uvitec Cambridge)* e respectivo *software*.

3.3 Preparação de controlo "mock"

Neste estudo, foi incluído um controlo "mock", constituído por DNA genómico de estirpes bacterianas puras de *Mycobacterium tuberculosis, Streptococus pneumoniae, Neisseria gonorrhoeae e Listeria monocytogenes*, que não são habitualmente encontradas no ambiente. Estas estirpes foram cedidas pelo Departamento de Doenças Infeciosas do INSA. O DNA destas estirpes (5 µl) foi corrido a uma concentração de 100ng/µl cada, em gel de agarose 0,8%, para avaliar a integridade do DNA genómico. Para cada estirpe foram feitos três replicados e obtidas três leituras de concentração para cada replicado usando o *Qubit®3.0*, com o kit BR. Após as quantificações, os replicados foram diluídos para a concentração de 1ng/µl e foi feito um *pool* com 2µl de cada replicado. Este *pool* final foi usado como controlo positivo do estudo.

3.4 Preparação de bibliotecas para sequenciação

3.4.1 Seleção de primers e amplificação do DNA por PCR

As bibliotecas para metagenómica foram construídas a partir de fragmentos de regiões de interesse do gene *16S rRNA*, e amplificadas através da técnica da PCR. Inicialmente, foi realizada uma reacção da PCR com um par de *primers* externos iniciadores (8F e 1541R), de forma a produzir um fragmento quase completo do gene *16S rRNA*, que serviu de molde para a segunda PCR (*nested*), dirigidos às regiões V3-V4, conforme **tabela II** e **figura 4**. Estes PCRs foram realizados em 5 séries introduzindo um controlo negativo em cada uma das séries.

Tabela II. Localização e sequências dos primers externos e internos do gene 16S rRNA usados neste estudo. As regiões sublinhadas indicam as sequências overhang específicas para construção de bibliotecas Illumina.

Nome do primer	Direção	Localização	Sequência oligonucleotídica
8F	*Forward*	Extremidade 5´ do gene 16S rRNA	5'AGAGTTTGATCCTGGCTCAG 3'
1541R	*Reverse*	Extremidade 3´ do gene 16S rRNA	5'AAGGAGGTGATCCAGCCGCA 3'
341F	*Forward*	Região VH3	5'TCGTCGGCAGCGTCAGATGTGTATAGACAGAAGCCACG GGNGGCWGCAG 3'
805R	*Reverse*	Região VH4	5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACT ACHVGGGTATCTAATCC 3'

A mistura para cada reacção da primeira PCR foi efetuada com 6,25 µl de *2X KAPA HiFi Hot Start ready mix*, 0,187 µl de primer 8F a 10 µM, 0,187 µl de primer 1541R a 10 µM, 5,376µl de água bidestilada e 0,5 µl de DNA a 100 ng/µl, num volume total de 12,5 µl. A preparação dos fragmentos da PCR *nested* para sequenciação foi efetuada de acordo com o protocolo de sequenciação metagenómica que se baseia na amplificação do gene 16S rRNA dos procariotas (”16S Metagenomic Sequencing Library Preparation Part#15044223 Rev.B”, Illumina, 2013).

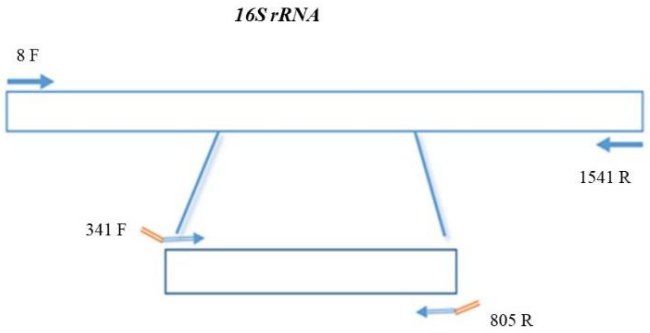


Figura 4. Representação simplificada do gene 16S rRNA, com localização dos primers externos e primers internos. As sequências overhang estão representadas por 2 linhas laranjas nos primers internos.

A reação da *nested* PCR foi preparada com 6,25µl de enzima *2x KAPA HiFi Hot Start ready mix*, 0,375µl de primer 341F, 0,375µl de primer 805R, 4,5µl de água bidestilada e 1 µl de produto resultante da primeira PCR, num volume total de 12,5 µl. Ambas as amplificações correram nas condições dos programas dos termocicladores de acordo com os esquemas representados na **figura 5**. Os produtos finais das PCR’s foram corridos em gel de agarose a 2% corado com brometo de etídeo, durante 1hora a 80 V e visualizados sob exposição de luz ultravioleta conforme indicado acima.

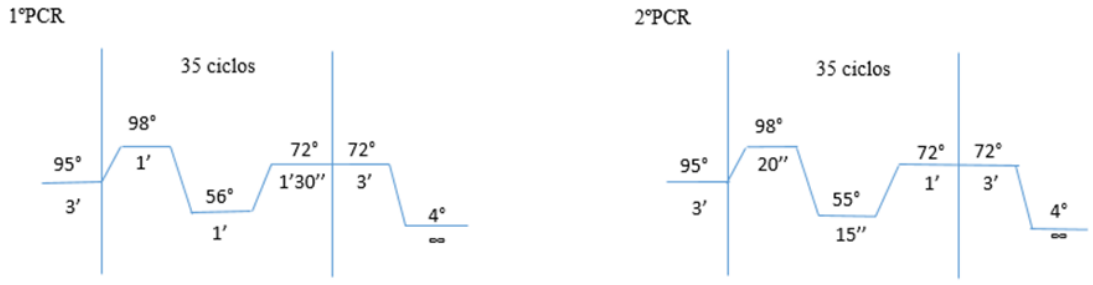


Figura 5. Representação esquemática dos programas da 1ª PCR e da nested PCR

3.4.2 Purificação dos produtos da PCR

Os dNTPs, *primers* e dímeros de *primers* da *nested* PCR foram removidos por purificação com *beads* que capturam o DNA e permitem limpar os restantes produtos da reação. Foi adicionado 20 µl de *AMPure XP beads* (*BeckmanCoulter*) em cada poço da microplaca contendo a amostra. Esta mistura foi gentilmente pipetada para homogeneizar as amostras e colocada a incubar durante 5 minutos à temperatura ambiente. A microplaca com as amostras foi colocada em suporte magnético durante 2 minutos até o sobrenadante ficar transparente. Neste momento as *beads* já capturaram o DNA e o sobrenadante foi descartado. O agregado *beads*/DNA foi então submetido a duas lavagens de 200 µl de etanol a 80% (sempre com a microplaca colocada no suporte magnético). Após as lavagens, o etanol foi eliminado, a microplaca foi removida do suporte magnético e as *beads* (contendo as amostras) foram deixadas ao ar durante 10 minutos até secarem por completo. As amostras foram ressuspendidas em 52,5 µl de 10mM Tris-HCL pH 8.5 e incubadas à temperatura ambiente por mais 2 minutos. A microplaca foi novamente colocada no suporte magnético durante 2 minutos e 50 µl do sobrenadante foram transferidos para uma nova microplaca de 96 poços.

3.4.3 PCR de indexação

Após a amplificação da região alvo, e da purificação dos produtos da PCR, as amostras foram duplicadas (com excepção de 2 amostras) de forma a totalizar 96 amostras em estudo, distribuídas da seguinte forma: 18 controlos negativos de extracção, 10 controlos negativos de PCR, 2 controlos positivos e 66 amostras.

Os produtos purificados foram submetidos a uma nova PCR para indexação das amostras, ficando cada amostra com um índice exclusivo. Cada índice tem uma sequência específica que será identificada pelo software do equipamento de sequenciação. Neste sentido, foi elaborada uma folha de trabalho com o esquema de pipetagem de forma a atribuir o índice correto a cada amostra e evitar trocas ou repetições de combinações de índices. A mistura da PCR foi preparada com 25 µl de 2x *kappa HiFi Hot start mix*, 5 µl de *Nextera XT Index primer 1*, 5 µl de *Nextera Index primer 2*, 10 µl de água para PCR e 10 µl de produto da PCR *nested*. Esta PCR correu com as seguintes condições: 95º - 3 minutos, 8 ciclos de 95ºC – 30 seg, 55ºC - 30seg e 72º C - 30seg, seguido de 72ºC - 5 minutos e incubação a 4ºC.

3.4.4 Quantificação, normalização e *pool* das bibliotecas

Para confirmação do tamanho esperado das bibliotecas, foi feito um gel de agarose a 2%, no qual se verificou que o tamanho era de ~630 pb, conforme exemplo da **figura 6**.

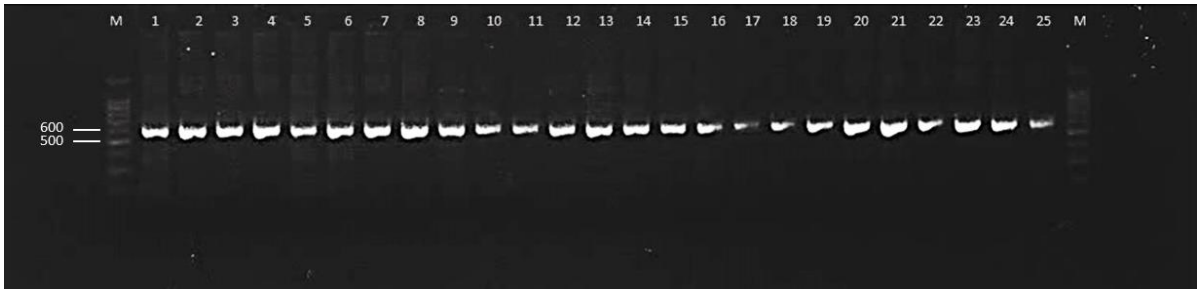


Figura 6. Exemplo de gel de agarose a 2% para 25 amostras do estudo, com o tamanho da biblioteca obtido de ~630 pb, onde M corresponde ao marcador de peso molecular de 100 pb.

As amostras foram quantificadas no *Qubit®3.0*, com o kit *High-sensitivity*, para determinação da concentração das bibliotecas em ng/µl. O cálculo da molaridade das bibliotecas foi efetuado com base no tamanho do produto da PCR (630 pb), e nas concentrações obtidas pelo *Qubit®3.0*, de acordo com a seguinte fórmula:

$$\frac{\text{(concentração em } ng/ul\text{)}}{\text{(660g/mol x tamanho da biblioteca)}} \times 10^6 = \text{concentração em nM}$$

As bibliotecas foram inicialmente diluídas para 30nM e depois para 4 nM usando o tampão de ressuspensão (Illumina). O *pool* final foi preparado usando 5 µl de cada biblioteca a 4nM

3.4.5 Desnaturação das bibliotecas e sequenciação

O *pool* de bibliotecas foi desnaturado (para que o DNA em cadeia simples se ligue aos adaptadores da *flowcell*) adicionando 5 µl de pool a 4 nM com 5 µl de solução NaOH 0,2 N, com incubação durante 5 minutos à temperatura ambiente. Após a desnaturação, os 10 µl foram diluídos em 990 µl de tampão de hibridação (HT1, Illumina) resultando numa diluição intermédia de 20 pM. Posteriormente, o DNA foi diluído para uma concentração final de 10 pM, juntando 300 µl da biblioteca a 20 pM com 300 µl de HT1. Foi também preparada uma biblioteca do genoma do bacteriófago PhiX de acordo com o procedimento descrito no guia *MiSeq System Denature and Dilute Libraries Guide* (Illumina, Doc# 15039740, v10, Fevereiro 2019), na mesma concentração de 10 pM. O objectivo da utilização do genoma de PhiX é de aumentar a diversidade das sequências nos *clusters* da *flowcell*, o qual foi adicionado numa proporção de 15% ao *pool* de bibliotecas. O *pool* e o PhiX foram carregados numa *cartridge* v3 de 600 ciclos e colocada no equipamento MiSeq (Illumina), que permite a sequenciação de 300 bases em cada extremidade dos fragmentos. A sequenciação consistiu em 2 *reads* de 301 pb para os fragmentos (sequenciação *paired-end*) e em 2 *reads* de 8 pb para as sequências de índices. O *workflow* utilizado foi o *Generate FastQ*, que efetua a geração de 2 ficheiros *fastq* de cada amostra, um respeitante à *read* 1 e outro à *read* 2 de cada fragmento, após desmultiplicação das sequências em cada *cluster* da *flowcell*. Estes ficheiros contêm as sequências e informação sobre a qualidade de cada nucleótido sequenciado, em formato *fastq.gz*.

4 Análise de dados

4.1 Análise primária

A desmultiplicação dos dados brutos (separação de *reads* por amostra de acordo com os índices utilizados), é efetuada automaticamente no equipamento MiSeq através da aplicação CASAVA 1.8, desenhada para fazer a conversão de ficheiros *.bcl* em ficheiros *.fastq*. A análise primária da corrida foi efetuada no software do equipamento, utilizando o *Sequencing Analysis Viewer 1.8.46*, onde se fez uma avaliação global dos dados gerados. Estes dados foram sumarizados num relatório utilizando o programa MultiQC V1.6. dev0[17], que também recorre ao programa FastQC [18] para obter diversas métricas de qualidade de cada par amostra/*read*.

4.2 Análise secundária

A análise secundária foi realizada com recurso a um software aberto, o *Quantitative Insights Into Microbial Ecology* "QIIME2", [13] desenhado para comparação e análise de comunidades microbianas a partir de dados gerados de NGS. Este software permite montar uma *pipeline* de análise recorrendo a ferramentas bioinformáticas que possibilitam filtragens de sequências mais finas, reduzir erros de sequenciação, fazer alinhamentos, inferir filogenias e criar árvores filogenéticas. Os ficheiros resultantes da desmultiplicação (*.fastq.gz*), produzidos pelo *workflow Generate FastQ* do equipamento MiSeq, foram transferidos para um servidor linux com 12 processadores, 32 Gb de memória e sistema operativo CentOS 6.8 onde se fez toda a análise com o programa QIIME2 em ambiente miniconda. Os ficheiros *fastq.gz* foram importados para o QIIME2 passando a ter uma extensão .qza, específica do software, de acordo com os seguintes comandos.

```
qiime tools import\
--type 'SampleData [PairedEndSequencesWithQuality]' \
--input-path /mnt/san/miseq/190709_M01600_0112_000000000-CCC94/Data/Intensities/BaseCalls \
--input-format CasavaOneEightSingleLanePerSampleDirFmt \
--out-put-path /home/dina.carpinteiro/Analise2/demux-paired-end.qza
```

Após a obtenção dos ficheiros .qza, foi aplicado o *plugin* DADA2. Este *plugin* permite fazer um *denoising* às amostras, ou seja, permite detetar e corrigir erros de sequenciação, reduzindo o "ruído" das sequências, como por exemplo remoção de regiões com sequências de baixa qualidade, e filtrar sequências quiméricas. A parametrização dos comandos para remoção de sequências com baixa qualidade, *p-trunc-len-f* e *p-trunc-len-r* foi efetuada de acordo com a observação feita nos gráficos de qualidade obtidos pelo *QIIME2* e pelo *MultiQC report* (**Anexo A)**, referindo-se às posições na qual as sequências devem ser truncadas na extremidade 3´, por diminuição da qualidade.

Neste trabalho foram testadas várias posições de truncagem no final das sequências, de modo a perceber qual o efeito no número de *reads* totais, assim como foram aplicados outros parâmetros como o *p-trim-left-f* e o *p-trim-left-r* para eliminar as bases correspondentes aos *primers* específicos do gene 16S rRNA no início de cada sequência. Outros parâmetros testados foram o *p-max-ee-f* e o *p-max-ee-r*. Estes parâmetros correm por defeito com um valor mínimo de 2 para ambas as *reads*, indicando que o número máximo de erros que aceita é de 2. Da aplicação destes comandos resultaram os ficheiros de saída que incluem uma *FeatureTable[frequency]* que indica o número de contagens (frequências) de sequências únicas em cada amostra e uma *FeatureData[Sequence]* com indicação das características para cada amostra, para cada teste efetuado.

```
qiime dada2 denoise-paired \
--i-demultiplexed-seqs/home/dina.carpinteiro/Analise2/demux-paired.qza
--p-trunc-len-f 290 \ 270; 260; 270
--p-trunc-len-r 270 \ 230; 230; 270
--p-trim-left-f 17 \
--p-trim-left-r 21 \
--p-max-ee-f 2 \ 3
--p-max-ee-r 5 \
--o-representative-sequences rep-seq-dada2.qza
--o-table table-dada2.qza\
--o-denoising-stats stats-dada2.qza \
--p-n-threads 2
```

Para visualizar os dados resultantes da filtragem de qualidade, os ficheiros .qza são transformados em ficheiros .qzv e visualizados na página view.qiime2.org.

```
qiimemetadatatabulate \
--m-input-file stats-dada2.qza \
--o-visualization stats-dada2.qzv

qiime feature-table tabulate-seqs \
--i-data rep-seqs-dada2.qza \
--o-visualization rep-seqs-dada2.qzv

qiime feature-table summarize \
--i-table table-dada2.qza \
--o-visualization table-dada2.qzv
```

A informação das contagens (frequências) de sequências únicas em cada amostra, contida no ficheiro *FeatureTable[frequency]*, requerem a aplicação de métricas para construir uma árvore filogenética que relacione as características das amostras entre si. Para a construção da árvore filogenética, foi utilizada uma *pipeline* de alinhamento das sequências com recurso às ferramentas *mafft*, *mask* e *fasttree* a partir do plugin *q2-phylogeny*. Primeiro, utilizou-se o programa *mafft* para executar um alinhamento de várias sequências do ficheiro *FeatureData[Sequence]* e criar um artefacto QIIME2 chamado *FeatureData[AlignedSequence]*.

```
qiime alignment maft \
--i-sequences/home/dina.carpinteiro/Analise9/Denoised/rep-seqs-dada2.qza \
--o-alignment /home/dina.carpinteiro/Analise9/aligned-rep.seqs.qza
```

Em seguida, foi utilizado o programa *mask* para remover posições altamente variáveis do alinhamento e que causam "ruído", não sendo informativas e até ambíguas para a construção da árvore filogenética.

```
qiime alignment mask \
--i-sequences /home/dina.carpinteiro/Analise9/Denoised/aligned-rep.seqs.qza \
--o-alignment /home/dina.carpinteiro/Analise9/masked-aligned-rep-seqs.qza
```

Após o alinhamento foi aplicado o programa *fasttree* para criar a árvore filogenética a partir do alinhamento efetuado. O programa *fasttree* cria uma árvore sem raíz. Com esta árvore sem raiz não é suposto fazer ainda qualquer inferência direta sobre o ancestral comum.

```
qiime phylogeny fasttree \
--i-alignment /home/dina.carpinteiro/Analise2/masked-aligned-rep-seqs.qza \
--o-tree /home/dina.carpinteiro/Analise2/unrooted-tree.qza
```

No final desta etapa é ainda aplicado o comando *midpoint tree*, cujo objetivo é calcular as distâncias entre duas extremidades e colocar a raíz no meio de dois pontos. Se a evolução for constante, esse ponto deve representar o ponto ancestral da árvore filogenética.

```
qiime phylogeny midpoint-root \
--i-tree /home/dina.carpinteiro/Analise9/unrooted-tree.qza \
--o-rooted-tree /home/dina.carpinteiro/Analise9/rooted-tree.qzv
```

Com a árvore filogenética enraizada, procederam-se às análises de diversidade, aplicando o método *core-metrics-phylogenetic*, que calcula várias métricas de diversidade alfa e beta e dos quais resultaram gráficos de análise de componentes principais (*Principal Component Analysis*, PCA).

```
qiime diversity core-metrics-phylogenetic \
--i-phylogeny /home/dina.carpinteiro/Analise9/rooted-tree.qza \
--i-table /home/dina.carpinteiro/Analise9/Denoised/table.dada2.qza \
--p-sampling-depth5000 \
--m-metadata-file /home/dina.carpinteiro/Analise9/metadata.tsv \
--output-dir /home/dina.carpinteiro/Analise9/core-metrics-results
```

As métricas principais resultantes do método *core-metrics-phylogenetic* foram:

Diversidade Alfa:
- Índice de diversidade *Shannon´s*
- Diversidade filogenética de *Faith´s*

Diversidade Beta:
- Distância de *Jaccard*
- Distância *Unweighted UniFrac*

Na parametrização do *–p-sampling-depth* foi usada a profundidade de 5000, considerando os dados da *table_dada2.qzv*, obtida após o *denoising*, que resume frequências e número de sequências representativas para cada amostra. Este parâmetro de profundidade corresponde à frequência total a que cada amostra deve ser rarefeita, para se obter uma amostragem uniforme. As métricas de diversidade beta foram também correlacionadas com o ficheiro de metadados (**Anexo B**), que contém a informação sobre as amostras, utilizando o método *diversity beta-group-significance* para determinar se existiam diferenças estatísticas significativas entre os grupos de amostras e controlos em estudo.

```
qiime diversity beta-group-significance \
--i-distance-matrix/home/dina.carpinteiro/Analise9/core-metrics-
results5000/unweighted_unifrac_distance_matrix.qza \
--m-metadata-file /home/dina.carpinteiro/Analise2/metadata.tsv \
--m-metadata-column Sample_Type \
--p-pairwise \
--o-visualization/home/dina.carpinteiro/Analise9/core-metrics-results5000/
unweighted_unifrac_Sample_Type_significance.qzv
```

```
qiime diversity beta-group-significance \
--i-distance-matrix/home/dina.carpinteiro/Analise9/core-metrics-results5000/jaccard_distance_matrix.qza \
--m-metadata-file /home/dina.carpinteiro/Analise2/metadata.tsv \
--m-metadata-column Sample_Type \
--p-pairwise \
--o-visualization/home/dina.carpinteiro/Analise9/core-metrics-results5000/jaccard_Sample_Type_significance.qzv
```

Para o estudo de significância da diversidade alfa, foi escolhida a métrica de diversidade filogenética de *Faith* [19]. Esta é uma medida qualitativa do enriquecimento de comunidades que incorpora relações filogenéticas entre características das populações, tendo sido efetuada com o seguinte comando:

```
qiime diversity alfa-group-significance \
--i-alpha-diversity /home/dina.carpinteiro/Analise9/core-metrics-results5000/faith_pd_vector.qza \
--m-metadata-file /home/dina.carpinteiro/Analise2/metadata.tsv \
--o-visualization /home/dina.carpinteiro/Analise9/core-metrics-results5000/faith_pd_significance.qzv
```

Após estas análises, foi efetuada a rarefação alfa para demonstrar o impacto nas amostras a diferentes profundidades. A parametrização de *–p-max-depth (100000)* teve a sua escolha também baseada na *table_dada2.qzv*.

```
qiime diversity alpha-rarefaction \
--i-table /home/dina.carpinteiro/Analise2/core-metrics-results/rarefied_table.qza \
--p-max-depth 100000\
--m-metadata-file /home/dina.carpinteiro/Analise2/metadata.tsv\
--p-steps 20\
--o-visualization /home/dina.carpinteiro/Analise2/core-metric-results/alpha_rarefaction_100000.qzv
```

Para explorar a composição taxonómica das amostras foi atribuído às sequências a taxonomia, através do *plugin feature-classifier*, que compara as sequências representativas de cada amostra com sequências de genomas de microrganismos de referência conhecidos. Para isso, os genomas foram importados através da base de dados de genomas de referência SILVA [20]. Os seguintes comandos foram utilizados:

```
qiime feature-classifier extract-reads \
 --i-sequences /home/dina.carpinteiro/Analise9/silva-138-99-seqs.qza \
 --p-f-primer AAGCCACGGGNGGCWGCAG \
 --p-r-primer GACTACHVGGGTATCTAATCC \
--p-min-length 445 \
 --p-max-length 485 \
--o-reads /home/dina.carpinteiro/Analise9/ref-seqs.qza

qiime feature-classifier fit-classifier-naive-bayes \
--i-reference-reads /home/dina.carpinteiro/Analise9/ref-seqs.qza \
 --i-reference-taxonomy /home/dina.carpinteiro/Analise9/silva-138-99-tax.qza \
 --o-classifier /home/dina.carpinteiro/Analise9/classifier.qza

qiime feature-classifier classify-sklearn \
--i-classifier /home/dina.carpinteiro/Analise9/classifier.qza \
 --i-reads /home/dina.carpinteiro/Analise9/rep-seqs-dada2.qza \
 --o-classification /home/dina.carpinteiro/Analise9/taxonomy.qza

qiime metadata tabulate \
 --m-input-file /home/dina.carpinteiro/Analise9/taxonomy.qza \
 --o-visualization /home/dina.carpinteiro/Analise9/taxonomy.qzv

qiime taxa barplot \
--i-table /home/dina.carpinteiro/Analise9/core-metrics-results-5000/rarefied_table.qza \
--i-taxonomy /home/dina.carpinteiro/Analise9/taxonomy.qza \
--m-metadata-file /home/dina.carpinteiro/Analise2/metadata.tsv \
--o-visualization /home/dina.carpinteiro/Analise9/taxa_bar_plots.qzv
```

Após a obtenção da taxonomia, efetuou-se uma filtragem para se obter sequencias que não se observassem em uma só amostra ou controlo, com o objectivo de reduzir o número de taxa total e facilitar as análises subsequentes nos vários níveis taxonómicos, através do seguinte comando:

```
qiime feature-table filter-features \
--i-table/home/dina.carpinteiro/Analise9/table-dada2.qza \
--p-min-samples 2 \
--o-filtered-table /home/dina.carpinteiro/Analise9/sample-contingency-filtered-table.qza
```

Após a obtenção do gráfico de barras com a taxonomia das amostras, foi utilizada a *pipeline* do Decontam [15], um pacote gratuito que funciona em R V3.4.2 (R Core Team, 2017), que implementa procedimentos de classificação estatística para identificação de contaminação em amostras de microbioma. No Rstudio V1.3.1073 (RStudio Team, 2020) foram ativados os pacotes necessários (phyloseq e qiime2R) e estabelecido o caminho para colocar os resultados finais.

```
setwd("C:/Users/dinac/Desktop/Analise9")
library(decontam)
library(phyloseq)
library(qiime2R)
```

Para iniciar as análises foi criado um ficheiro *physeq* que correlacionou a informação proveniente do *QIIME2*, nomeadamente a tabela de características *feature-table.qza*, e o ficheiro de metadados *metadata.tsv* e o ficheiro correspondente com a taxonomia, *taxonomy.qza:*

```
physeq<- qza_to_phyloseq(features="sample-contingency-filtered-table.qza" , metadata = "metadata.tsv"
, taxonomy = "taxonomy.qza" )
```

Depois, foi atribuído aos controlos de extração a informação que seriam controlos negativos, para poder prosseguir com as análises seguintes:

```
sample_data(physeq)$is.neg<- sample_data(physeq)$Sample_Type == "Extraction_Control"
```

A primeira função usada no Decontam foi o *isContaminant*, que usa padrões baseados na prevalência (presença ou ausência) entre as amostras e controlos para identificar sequencias contaminantes. Assim, uma vez que já se tinha definido que o controlo da extração seria um controlo negativo, o *script* seguinte verificou qual a prevalência de ASV´S contaminantes presentes nas amostras.

```
contamdf.prev_ExtractionControl<- isContaminant(physeq, method="prevalence", neg="is.neg")
table(contamdf.prev_ExtractionControl$contaminant)
```

O mesmo processo foi efetuado para os outros controlos negativos introduzidos no estudo, correspondentes aos controlos da PCR.

```
sample_data(physeq)$is.neg<- sample_data(physeq)$Sample_Type == "PCR_Control"
```

A caracterização dos controlos da PCR também como controlos negativos, permitiu que o *script* seguinte verificasse qual a prevalência de ASV´S contaminantes presentes nas amostras.

```
contamdf.prev_PCR_Control<- isContaminant(physeq, method="prevalence", neg="is.neg")
table(contamdf.prev_PCR_Control$contaminant)
```

Uma vez que os contaminantes já foram identificados, estes foram removidos do objecto *physeq* através dos seguintes *scripts* e serão movidos para outros objetos, passando a ter-se a informação individualizada dos taxa contaminantes e não contaminantes.

```
#mover para outro objeto os taxa não contaminantes

physeq.noncontam_ExtractionControl<-prune_taxa(!contamdf.prev_ExtractionControl$contaminant, physeq)

physeq.noncontam_PCR_Control<- prune_taxa(!contamdf.prev_PCR_Control$contaminant, physeq)


#mover para outro objeto os taxa contaminantes

physeq.contam_ExtractionControl<-prune_taxa(contamdf.prev_ExtractionControl$contaminant, physeq)

physeq.contam_PCR_Control<- prune_taxa(contamdf.prev_PCR_Control$contaminant, physeq)
```

Os gráficos de barras dos filos e géneros contaminantes foram obtidos com os seguintes *scripts*:

```
plot_bar(physeq.contam_ExtractionControl , fill="Genus")

plot_bar(physeq.contam_ExtractionControl , fill="Phylum")

plot_bar(physeq.contam_PCR_Control, fill="Genus")

plot_bar(physeq.contam_PCR_Control, fill="Phylum")
```

Para determinar a frequência dos diferentes contaminantes e a sua proporção relativa nas amostras, relacionaram-se os dados totais de sequências obtidas após o *denoising* com a frequência de taxa contaminantes obtidos para cada amostra. Estes dados foram trabalhados em folha de cálculo do *Microsoft Excel.*

5 Resultados

5.1 Análise de qualidade das sequências

As métricas gerais de qualidade dos dados resultantes da sequenciação, foram analisadas através do *MultiQC* **(Anexo A)**. O rendimento da sequenciação foi avaliado com base no número total de *reads* obtidas e na qualidade de bases sequenciadas com índice de qualidade Q30 ou superior, tendo em conta as especificações do kit Illumina utilizado (Kit MiSeq V3-600 ciclos) (**tabela III**). Após a sequenciação obteve-se um total aproximado de 47,3 milhões de *reads*, que correspondem a um total de 12,6 Gb de sequência (especificações MiSeq-Illumina: 13,2-15 Gb de sequência total). A qualidade global da corrida (% de bases >=Q30) foi de 68%, o que é inferior à especificação do fabricante (% Q30 >= 70%), mas que é aceitável tendo em conta que a sequenciação de amplicões do gene 16S rRNA tem subjacente uma baixa diversidade, que afecta a normalização dos sinais de fluorescência das 4 bases e, subsequentemente, a qualidade do respectivo *base-calling*. A taxa de alinhamento do PhiX foi de 20%, estando ligeiramente acima do esperado (15%), devendo-se esta alteração, provavelmente a pequenas variações nas quantificações da biblioteca e do PhiX. A taxa de erro da sequenciação do PhiX variou entre 2,1% na *read* 1 e 2,5% na *read* 4.

Tabela III. Sumário de métricas gerais de qualidade do MultiQC report.

Read	***M Reads***	***Rendimento Mb Bp***	***Alinhamento do PhiX (%)***	***Taxa de erro do PhiX (%)***	***% >= Q30***
Read 1	23,65	6.170,0	20,1%	2,1%	72,1%
Read 2 (I)	23,65	140,0	0,0%	nan%	68,8%
Read 3 (I)	23,65	140,0	0,0%	nan%	84,9%
Read 4	23,65	6.170,0	20,0%	2,5%	63,5%
Total	47,3	12.630,0	20,0%	2,3%	68,0%

No *QIIME2*, obtiveram-se os gráficos de qualidade das *reads* (**figura 7**) através do ficheiro *demux-paired-end.qzv*, a partir do qual se observaram os perfis de qualidade para as *reads forward* e *reverse*. Os perfis obtidos mostram que ambas as *reads* apresentam um decréscimo de qualidade (abaixo de Q20) a partir das 220 pb na *forward read* e a partir das 200 pb na *reverse read*. No geral, a *forward read* apresenta melhor qualidade em comparação com a *reverse read.*

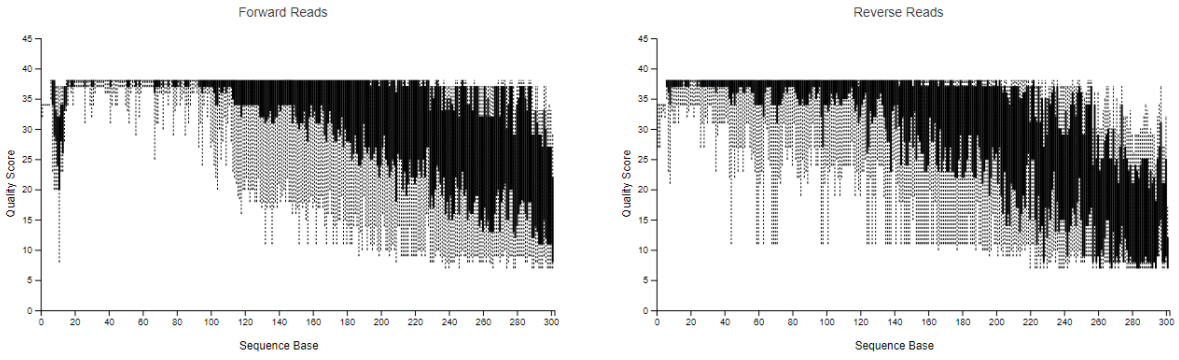


Figura 7. Representação gráfica da qualidade para as forward reads e reverse reads, obtida através do QIIME2.

5.2 Tratamento de sequências

O processo de sequenciação de amplicões em plataformas de NGS introduz erros nos dados de sequenciação, podendo esses erros levar a dificuldades de interpretação de resultados [14]. Neste sentido, e com base na informação dos gráficos gerados, foi efetuado o *denoising* às amostras com o DADA2 para remoção de sequências com baixa qualidade (**filtradas**), correcção de erros de sequenciação (***denoised***) e remoção de quimeras (**não quiméricas**), testando 9 condições distintas (análises 1 a 9) com diferentes parâmetros de truncagem, *trimming*, e número de erros máximo aceitável, cujos resultados se encontram resumidos na **tabela IV**. Durante este processo, a perda de *reads* apenas acontece nas etapas de filtragem, fusão das sequências emparelhadas (*merge*) e identificação de sequências quiméricas, enquanto que na etapa *denoising* é atribuído a todas as *reads* uma sequência corrigida, inferida pelo próprio modelo DADA2.

Nas primeiras 5 análises fizeram-se variar apenas os parâmetros de truncagem das *reads*, de forma a observar como estas alterações afectavam os resultados finais. De uma forma geral, os resultados da truncagem mostraram que não houve grandes variações entre as análises no número de *reads* finais (variação de 4-7%). A maioria das perdas de *reads* observou-se logo na filtragem. Na análise 5 onde a extensão de corte foi maior, deixando somente um *overlap* máximo entre a *forward read* e a *reverse read* de ~20 bases (*p-trunc-len-f 250* e *p-trunc-len-r 230*), obteve-se mais sequências na filtragem (50%), ainda que no final o número de *reads* apenas representasse 7% do total (9767778 *reads*). Em todas as análises, entre o denoising e o merge, não há perdas significativas de *reads* (variação de 4-7%), mantendo-se o número de *reads* acima dos 90%. Nas não-quiméricas voltam-se a perder muitas *reads*, mantendo-se apenas entre os 14-16%. Após estes testes iniciais, foram acrescentados na análise 6 os parâmetros para remoção dos *primers* usados na amplificação do gene *16S rRNA*, com a seguinte parametrização: *p-trim-left-f 17* e *p-trim-left-r 21*. Os resultados mostraram que apesar de se terem perdido novamente mais *reads* na filtragem inicial, uma vez que só se reteve 31% das *reads* iniciais, ganhou-se mais *reads* no final, em resultado de se perderem muito menos *reads* na análise de quimeras, resultando no final em 22% de *reads* (cerca de três vezes mais que o obtido na análise 5). O *merge* das sequências continuou acima dos 90%, indicando a eficácia na junção da maioria das *forward reads* com as *reverse reads.*

O parâmetro *p-max-ee-f* corre por base na *pipeline* de análise do DADA2 com valores de 2 para ambas as *reads*. Este parâmetro indica o número máximo de erros esperados, removendo todas as sequências que tenham um número de erros acima deste valor. Na análise 7 foram testados os valores *p-max-ee-f 3* e *p-max-ee-r* 5, e na análise 8 os valores *p-max-ee-f 2* e *p-max-ee-r* 5. Os resultados mostraram que, na filtragem, a análise 7 obteve mais *reads* (56%) do que na análise 8 (49%) e, consequentemente, mais *reads* no final, em resultado de se ter permitido um maior número de erros nas *forward reads.*

Com este conjunto de resultados, fez-se então a última análise, juntando os parâmetros que obtiveram melhores resultados. Desta análise 9, obteve-se o maior número de *reads* na filtragem (68%), o *merge* obteve valores acima dos 90%, continuando a manter a junção da maioria das *forward* e *reverse reads*, e o resultado final manteve 48% das *reads* em relação às *reads* iniciais. Os resultados desta última análise (~6,87M de *reads*) foram usados para dar continuidade às análises de filogenia.

Tabela IV. Resumo das análises efetuadas com o plugin DADA2 e respetivos resultados, indicando o número de reads iniciais e o número de reads finais. Os resultados estão representados como uma percentagem do número de reads em cada passo relativamente ao número de reads do passo anterior, excepto na "% reads finais" em que a percentagem respeita ao número de reads final relativamente ao número de reads inicial.

	Análises								
Parâmetros	**Análise 1**	**Análise 2**	**Análise 3**	**Análise 4**	**Análise 5**	**Análise 6**	**Análise 7**	**Análise 8**	**Análise 9**
P-trunc-len-f	290	270	260	270	250	270	270	270	250
P-trunc-len-r	270	230	230	270	230	270	270	270	230
P-max-ee-f	2	2	2	2	2	2	3	2	3
P-max-ee-r	2	2	2	2	2	2	5	5	5
P-trim-left-f	0	0	0	0	0	17	0	0	17
P-trim-left-r	0	0	0	0	0	21	0	0	21
	Resultados								
Nº reads iniciais	14277243	14277243	14277243	14277243	14277243	14277243	14277243	14277243	14277243
Filtradas	36%	45%	48%	30%	50%	31%	56%	49%	68%
Denoised	98%	98%	98%	98%	99%	99%	98%	98%	99%
Merged	94%	94%	94%	92%	94%	94%	91%	92%	95%
Não-quiméricas	16%	15%	14%	16%	15%	79%	17%	16%	76%
% reads finais	5%	6%	6%	4%	7%	22%	8%	7%	48%
Nº reads finais	765282	901066	916377	640343	976778	3210971	1179752	1020876	6876905

5.3 Análises de Diversidade

Após o *denoising,* efetuaram-se análises de riqueza e da diversidade de sequências a nível de cada grupo e entre grupos. Estas análises permitiram avaliar se existiam diferenças na composição entre o grupo das amostras e os grupos dos controlos negativos adicionados ao estudo.

As análises possibilitaram também avaliar se o conjunto total de dados de sequenciação produzidos era suficiente para os objetivos deste trabalho.

Para isso foi inicialmente estabelecido um valor de profundidade para as amostras (*p-sampling-depth*). A profundidade escolhida foi de 5000 sequências por ser um valor abaixo do valor obtido para a amostra menos representada na tabela de frequências (22945_dup: 13043 sequências). A escolha desta profundidade de sequenciação teve em conta um valor considerado suficientemente alto para capturar a diversidade presente nas amostras, mas baixo o suficiente para não eliminar nenhuma amostra do estudo. Este método, *core-metrics-phylogenetic,* extrai sequências ao acaso de cada amostra (sem substituição) para que todas tenham a mesma profundidade[21].

5.3.1 Diversidade alfa

Para responder às questões "Quantas espécies diferentes podem ser detetadas numa determinada amostra ambiental?" e "Qual é a abundância relativa de cada espécie nessa amostra?", foram usados os índices de diversidade alfa, uma vez que estes reflectem a riqueza e diversidade de espécies presentes. No presente caso, a riqueza é representada pelo número de ASVs, ou seja, o número total de sequências distintas existente numa determinada amostra, podendo estas sequências constituir, à partida, diferentes espécies.

Efetuaram-se então testes de significância sobre as matrizes alfa obtidas, usando o teste de Kruskal-Wallis para determinar se havia diferenças estatisticamente significativas entre os grupos em estudo (amostras, controlos negativos e controlos positivos). De acordo com o *índice de Faith*, que avalia a riqueza dos grupos taxonómicos, os resultados obtidos mostraram que existe uma grande variação deste índice no grupo das amostras, ao contrário do que acontece nos grupos dos controlos (**figura 8**). Estes resultados sugerem que as amostras possuem *a priori* um conjunto de sequências que não se encontram nos controlos utilizados durante a sua manipulação.

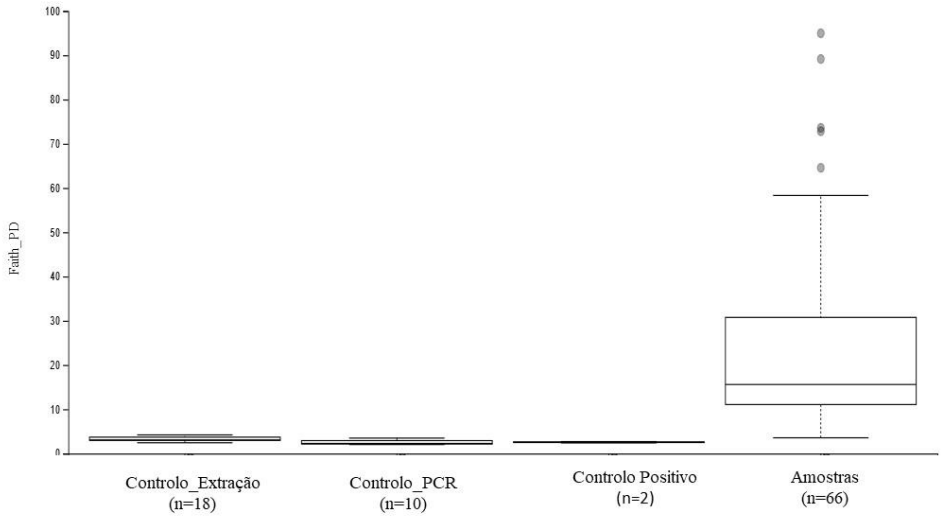


Figura 8. Caixas de bigodes do índice de Faith de acordo com os grupos em estudo.

O resultado do teste não paramétrico de Kruskal-Wallis mostrou que existem diferenças significativas entre todos os grupos para esta medida de riqueza (*p-value*<0.05). Também o teste estatístico de Kruskal-Wallis por pares (**tabela V**) confirmou a existência de diferenças estatisticamente significativas entre cada par de grupos em comparação, à excepção entre o controlo da PCR e controlos positivos e entre os controlos da extração e os controlos positivos, refletindo provavelmente o facto estes grupos terem um pequeno número de elementos.

Tabela V. Resultados do teste de Kruskal-Wallis por pares, para o índice de Faith, aplicado aos grupos em estudo.

Grupo 1	**Grupo 2**	**H**	**P-value**
Controlo Extração (n=18)	Controlo PCR (n=10)	8.554023	3.44759e-03
	Controlo Positivo (n=2)	3.571429	5.878172e-02
	Amostras (n=66)	40.252406	2.231803e-10
Controlo PCR (n=10)	Controlo Positivo (n=2)	0.184615	6.674365e-01
	Amostras (n=66)	25.714286	3.958857e-07
Controlo Positivo (n=2)	Amostras (n=66)	5.739130	1.659100e-02

A diversidade alfa também pode ser representada pelo índice de *Shannon* (também conhecido como índice de diversidade de *Shannon*), uma vez que este índice mede a diversidade de espécies numa determinada amostra, isto é, avalia se as diferentes espécies têm abundâncias equivalentes ou se uma ou mais espécies são mais abundantes que as restantes [22] [23]. De acordo com a **figura 9**, existe uma grande variação na diversidade no grupo das amostras, seguido dos controlos negativos das extrações, controlos da PCR e controlo positivo. Estes resultados sugerem que a diversidade de espécies contaminantes é maior a nível dos controlos negativos da extracção do que a nível dos controlos da PCR.

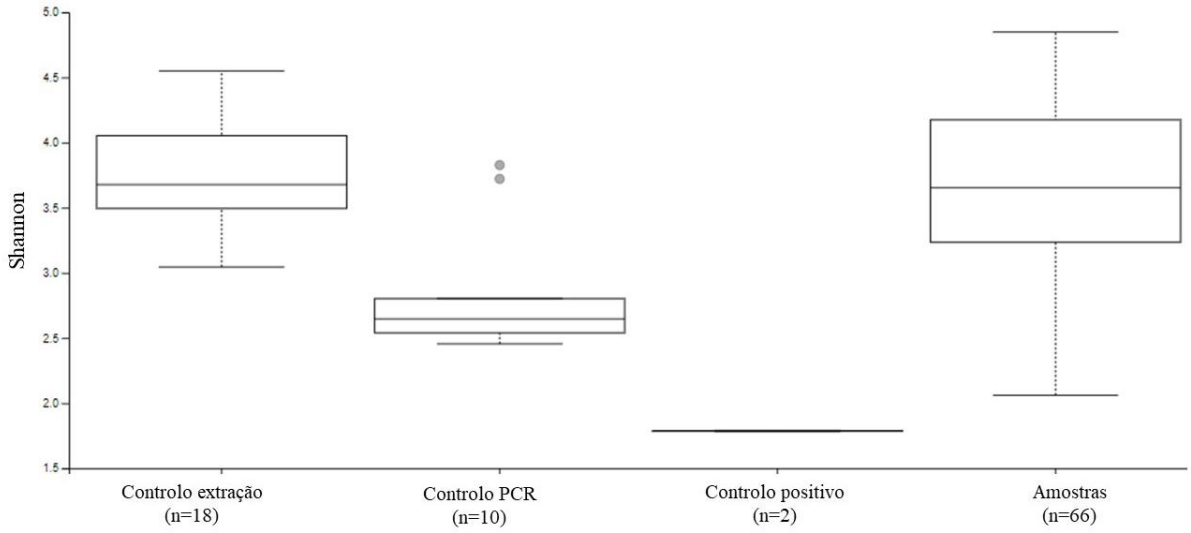


Figura 9. Caixas de bigodes para o índice de Shannon, de acordo com os grupos em estudo.

Quando visualizado o resultado do teste não paramétrico de Kruskal-Wallis, verifica-se que existem diferenças estatisticamente significativas entre todos os grupos, para esta medida de diversidade ($p<0.05$). No entanto, o teste estatístico de Kruskal-Wallis por pares (**tabela VI**) mostrou que não existe uma diferença significativa de diversidade entre os controlos da extracção e as amostras, sugerindo a existência de alguma complexidade a nível da composição microbiana dos reagentes, soluções e/ou consumíveis, usados na extracção de DNA. Em resumo, as análises de diversidade alfa mostraram que existem diferenças claras na composição microbiana entre o grupo das amostras e os grupos dos controlos negativos.

Tabela VI. Resultados do teste de Kruskal-Wallis por pares, para a diversidade de Shannon, aplicado aos grupos em estudo.

Grupo 1	Grupo 2	H	P-value
Controlo Extração (n=18)	Controlo PCR (n=10)	10.944828	9.39e-04
	Controlo Positivo (n=2)	5.142857	2.3342e-02
	Amostras (n=66)	0.162686	6.86696e-01
Controlo PCR (n=10)	Controlo Positivo (n=2)	4.615385	3.1686e-02
	Amostras (n=66)	10.215821	1.392e-3
Controlo Positivo (n=2)	Amostras (n=66)	5.739130	1.6591e-02

5.3.2 Rarefação alfa

Após a análise de métricas de diversidade alfa, foram geradas curvas de rarefação alfa para verificar se a amostragem realizada na profundidade de sequências escolhida (*sampling depth* de 5000), foi a apropriada para capturar toda a diversidade contida nos controlos (controlos da extração e controlos da PCR). Para gerar as curvas de rarefação alfa, utilizaram-se níveis de profundidade entre 5000 e 100000 (com amostragem em 20 intervalos). O gráfico obtido para o índice de *Shannon* parece demonstrar que a profundidade escolhida foi adequada para os controlos **(figura 10)**. As linhas dos grupos dos controlos atingem um platô, indicando que a diversidade nestes grupos foi totalmente capturada com 5000 sequências e que não se torna necessária uma maior profundidade de sequenciação para obter toda a diversidade existente. No entanto, caso o objectivo fosse capturar todas as estirpes existentes no grupo das amostras, a profundidade de sequenciação teria de ser no mínimo de 80000.

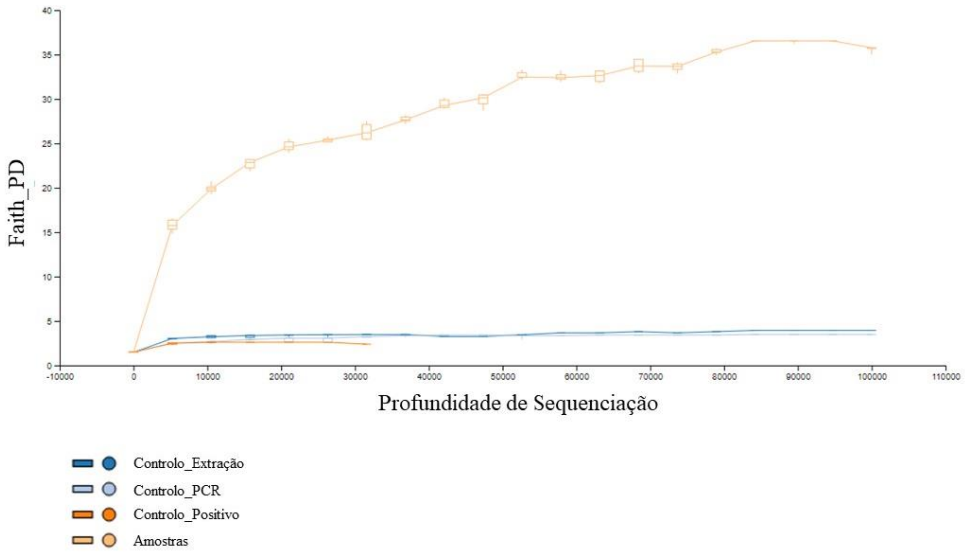


Figura 10. Gráfico da rarefação alfa que mostra a diversidade de Faith para cada grupo em função da profundidade de sequenciação.

5.3.2 Diversidade beta

A diversidade beta procura quantificar a diferença na composição de espécies entre 2 ambientes microbianos ou amostras, usando os perfis de abundância dos vários taxa. Neste trabalho, foram utilizadas 2 medidas de diversidade beta que se baseiam em diferentes tipos de dados, para caracterizar as diferenças entre grupos, nomeadamente, a distância de *Jaccard* e o método *UniFrac*.

A distância de *Jaccard* baseia-se na presença ou ausência de espécies entre 2 amostras, não tendo em consideração os dados de abundância [24]. Nesta medida, o valor de 0 (zero) significa que ambas as amostras têm exatamente as mesmas espécies, enquanto que o valor de 1 significa que não existem quaisquer espécies em comum entre essas amostras. Quanto mais semelhantes 2 amostras forem entre si, menor é a "distância" entre elas, ou seja, a distância tende para zero quando as amostras tendem a ter mais espécies em comum. Na **figura 11**, pode-se observar que as distâncias entre as amostras e os diversos controlos estão próximas do valor de 1, indicando que existem menos espécies em comum entre os grupos das amostras e dos controlos, do que dentro do grupo das amostras. É de referir, no entanto, que uma grande parte das amostras contem 2 replicados, sendo por isso de esperar que a distância entre estes resulte num valor baixo da distância de *Jaccard*.

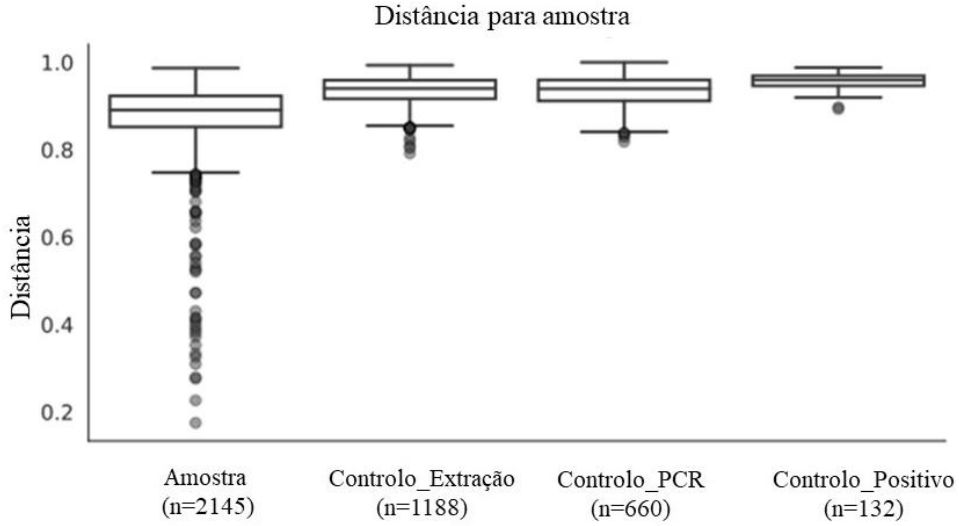


Figura 11. Caixas de bigodes com representação da distância de Jaccard entre cada amostra e os restantes membros de cada grupo do estudo (amostras, controlos Extração, Controlos PCR, e controlo positivo).

De acordo com o teste de Permanova, verifica-se a existência de uma diferença estatisticamente significativa entre os 4 grupos em estudo ($p<0.05$) (**tabela VII**). O teste de Kruskal-Wallis corrobora os valores observados no gráfico obtido, pois todas as associações entre grupos obtêm valores de *p-value* < 0.05 (**tabela VIII)**, com excepção da relação entre os grupos do controlo da PCR e do controlo positivo (*p-value* 0.066), que pode refletir o reduzido tamanho da amostra (n=12).

Tabela VII. Resultado do teste de Permanova para a distância de Jaccard.

	Resultados Permanova
Teste estatístico	Pseudo-F
Número de amostras	96
Número de Grupos	4
Teste estatístico	3.48884
P-value	0.001
Número de permutações	999

Tabela VIII. Resultado do teste de Kruskal-Wallis para a distância de Jaccard entre grupos.

Grupo 1	**Grupo 2**	**Tamanho amostra**	**P-value**
Controlo Extração	Controlo PCR	28	0.001
	Controlo Positivo	20	0.025
	Amostras	84	0.001
Controlo PCR	Controlo Positivo	12	0.066
	Amostras	76	0.001
Controlo Positivo	Amostras	68	0.002

Com base na distância de *Jaccard*, foi efectuada uma análise PCA que permitiu evidenciar uma boa separação dos grupos dos controlos de extração, controlos da PCR e controlos positivos, em relação ao grupo das amostras (**figura 12**). Conforme esperado em função dos resultados obtidos anteriormente, o grupo das amostras evidencia uma grande dispersão espacial, enquanto que os grupos dos controlos se encontram bastante mais compactos.

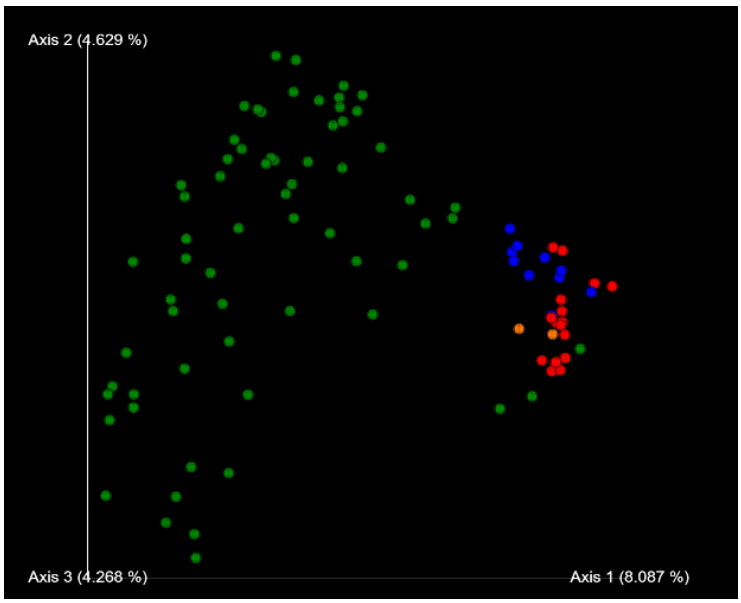


Figura 12. Representação da análise PCA para a distância de Jaccard, onde a cor verde corresponde às amostras, laranja ao controlo positivo, vermelha aos controlos de extração e azul aos controlos da PCR.

O método *UniFrac* é uma outra medida da diversidade beta que incorpora informação sobre a relação de parentesco dos membros das amostras [25]. Neste método, utilizam-se os valores de distância entre sequências calculados de acordo com a respectiva árvore filogenética. A distância entre as sequências é baseada na fracção do comprimento do ramo da árvore que é partilhado entre as amostras ou que é exclusivo de uma ou de outra amostra. O método *UniFrac* tem 2 sub-métodos distintos, *unweighted* e *weighted.* Neste trabalho foi usada a distância de *unweighted UniFrac*. Esta é uma medida qualitativa, que se baseia unicamente na distância entre sequências (presença ou ausência de ASVs) e não tem em consideração os dados de abundância (como acontece com o *weighted UniFrac*). Esta medida suporta os resultados obtidos para a distância de *Jaccard*, isto é, que existe uma maior distância entre ASVs dentro do grupo das amostras do que entre as amostras e os membros dos grupos controlos (**figura 13**).

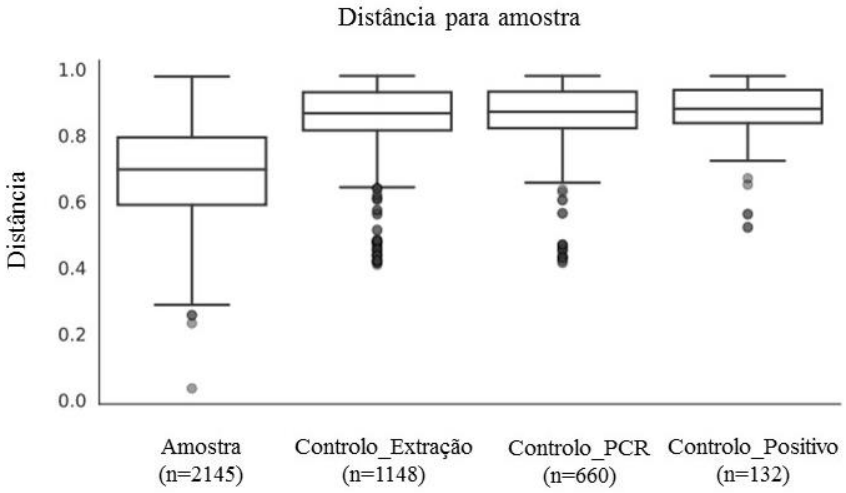


Figura 13. Caixas de bigodes com representação da distância Unweighted Unifrac entre cada amostra e restantes membros de cada grupo do estudo (amostras, controlos extração, controlos da PCR e controlo positivo).

Para avaliar as diferenças entre os diferentes grupos, foi aplicado o teste Permanova para a distância *unweighted Unifrac,* verificando-se que existem diferenças estatísticas significativas (*p-value*<0.05) (**tabela IX**). O teste de Kruskal-Wallis confirmou as diferenças estatisticamente significativas para a comparação entre cada par de grupos (**tabela X)**. Em resumo as medidas de diversidade beta mostraram existir diferenças significativas entre amostras e grupos controlo, no que respeita à composição taxonómica. Este tipo de análises são assim úteis para determinar se grupos de amostras de microbioma têm um perfil taxonómico similar ou distinto do perfil de grupos que contêm taxa contaminantes.

Tabela IX Resultados do teste Permanova aplicado para a medida de distância Unweighted Unifrac.

	Resultados Permanova
Teste estatístico	Pseudo-F
Número de amostras	96
Número de Grupos	4
Teste estatístico	14.7404
P-value	0.001
Número de permutações	999

Tabela X. Resultado do teste Kruskal-Wallis para a distância Unweighted Unifrac entre grupos.

Grupo 1	*Grupo 2*	*Tamanho amostra*	*P-value*
Controlo Extração	Controlo PCR	28	0.001
	Controlo Positivo	20	0.031
	Amostras	84	0.001
Controlo PCR	Controlo Positivo	12	0.034
	Amostras	76	0.001
Controlo Positivo	Amostras	68	0.001

5.4 Composição taxonómica

A composição taxonómica foi obtida através da correlação da tabela de frequências, que reúne a informação das frequências relativas das sequências obtidas por amostra, com a base de dados SILVA. A visualização da taxonomia foi efectuada através de gráficos de barras para diferentes níveis taxonómicos, conforme ilustrado para o nível de filo na **figura 14**.

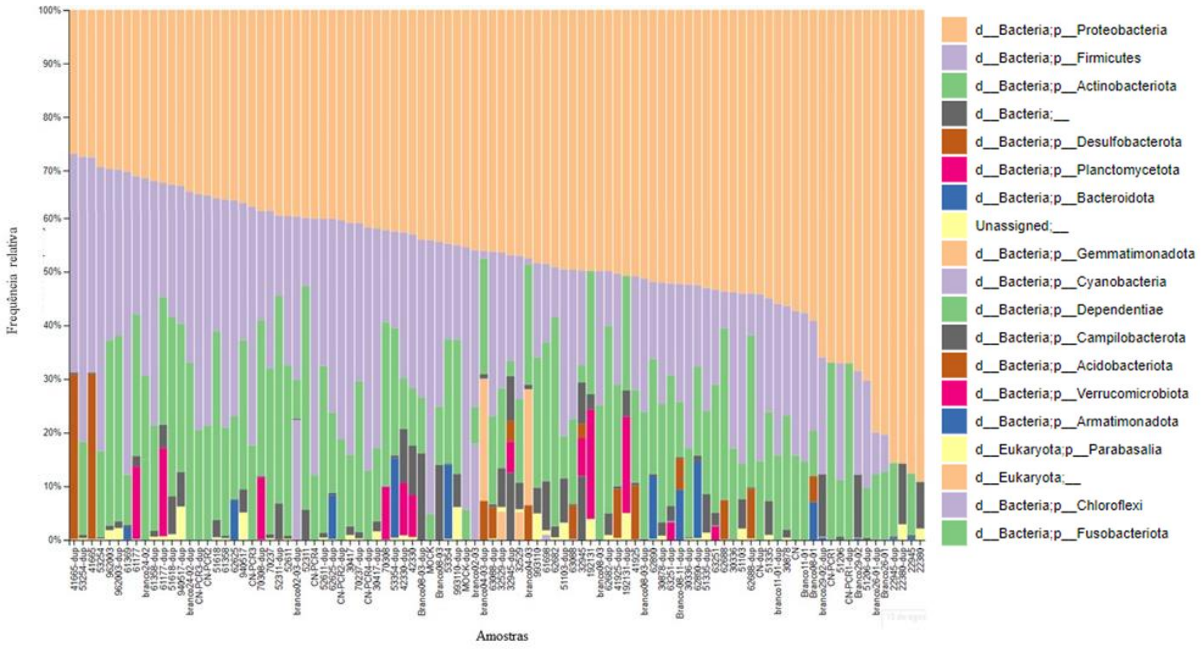


Figura 14. Composição taxonómica ao nível de filo, permitindo visualizar os filos mais abundantes e menos abundantes em cada amostra. A legenda dos filos encontra-se ordenada do mais frequente para o menos frequente

Para este nível taxonómico, observou-se que o filo *Proteobacteria* está presente em todas as 96 amostras do estudo. Os filos *Firmicutes* e *Actinobacteriota* são também muito comuns no conjunto total das amostras estudadas. No entanto, existem filos muito menos frequentes, como por exemplo os *Fusobacteriota* e *Chloroflexi*. Dado que este tipo de gráfico é interactivo na sua forma original, permitiu fazer uma seleção individual de cada filo, pelo que se observou que os filos *Cyanobacteria,*

Dependentiae, *Acidobacteriota* e *Armatimonadota* apenas se encontravam em controlos negativos da extração, indicando provável contaminação ambiental conforme já reportado em outros estudos [26].

Ao nível da composição taxonómica de espécies, observou-se a presença de 114 espécies diferentes no total das 96 amostras, o que dificultou a interpretação dos resultados (**figura 15**). No entanto, observou-se resumidamente que as amostras e seus respetivos duplicados aparentam partilhar as mesmas espécies entre si, ainda que com pequena variação nas proporções relativas, e algumas espécies parecem ser dominantes ao longo das amostras. Uma vez que o número de espécies com classificação taxonómica foi elevado, optou-se por prosseguir a análise dos contaminantes ao nível mínimo de género.

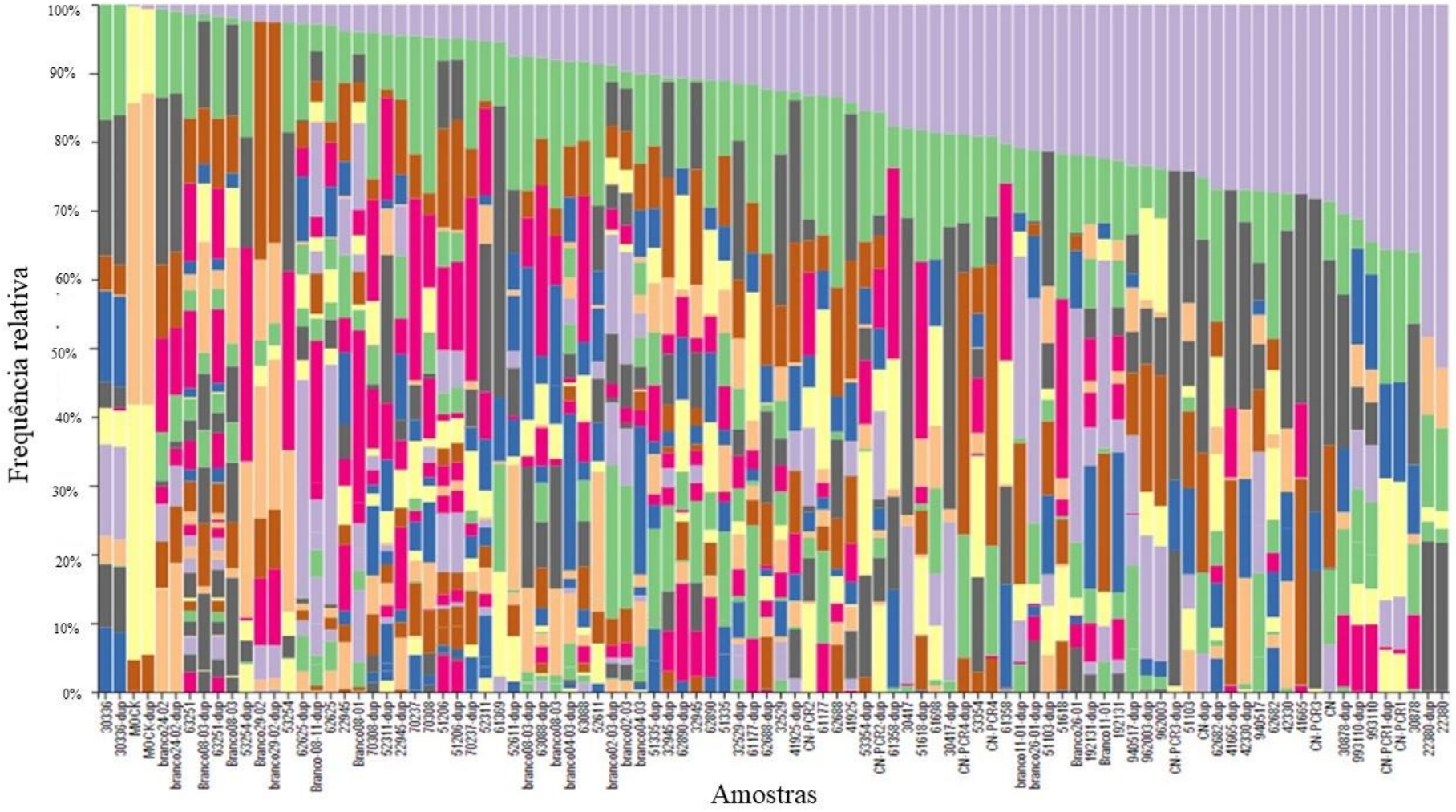


Figura 15. Composição taxonómica ao nível da espécie nas 96 amostras estudadas, nas quais foram identificadas 114 espécies diferentes em todos os amplicões sequenciados, A legenda da figura encontra-se no Anexo C.

5.4.1 Identificação dos contaminantes dos controlos de extração

Com o objetivo de analisar a composição taxonómica dos contaminantes presentes nos controlos negativos da extração e da PCR, e analisar o seu impacto nas amostras, utilizou-se o programa Decontam em ambiente R. Este programa permitiu identificar como contaminantes os taxa presentes nos controlos negativos (extracção e PCR) e filtrá-los das amostras com base na prevalência (presença/ausência) destes contaminantes. Foram assim gerados ficheiros individuais com a informação dos contaminantes presentes nos controlos das extrações e nos controlos da PCR. Destes ficheiros, obtiveram-se gráficos de barras que permitiram observar a composição taxonómica e abundância dos contaminantes ao nível de filo e de género. Nos controlos da extração, ao nível de filo, identificaram-se 10 filos contaminantes, sendo o filo *Proteobacteria* o mais dominante e abundante, sendo consequentemente também o filo contaminante mais presente e abundante nas amostras (**figura 16**).

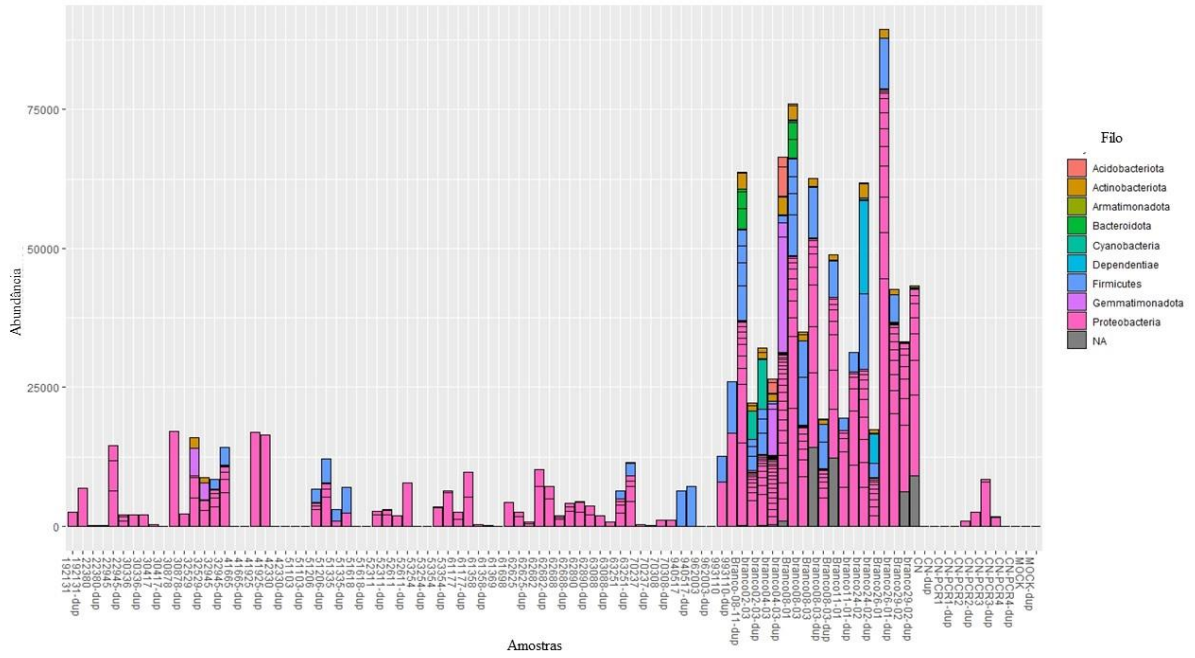


Figura 16. Abundância dos filos contaminantes dos controlos da extração, nos 4 grupos em estudo.

No grupo das amostras, apenas 8 das 66 amostras, não evidenciaram qualquer tipo de contaminação proveniente dos controlos da extração. As proteobactérias estão presentes em 55 amostras e, na maior parte destas, este é o único filo contaminante, conforme se verifica na **tabela XI**. O filo *Firmicutes* também é um contaminante frequente, que foi identificado em 14 amostras, com uma proporção a variar entre 20-100%. Para além destes contaminantes, os filos *Actinobacteriota, Gemmatimonadota e Bacteroidota* obtiveram uma frequência igual ou inferior a 3 amostras, enquanto que os filos *Acidobacteriota, Armatimonadota, Cyanobacteria* e *Dependentiae* identificados nos controlos negativos da extracção, não foram identificados em nenhuma amostra.

Tabela XI. Frequência e proporção relativa de filos contaminantes (dos controlos da extração) presentes nas amostras.

Filos Contaminantes	Frequência	Mínimo	Máximo
Proteobacteria	55	0,2%	100%
Firmicutes	14	20%	100%
Actinobacteriota	3	0,1%	12%
*NA**	3	0,2%	0,3%
Gemmatimonadota	2	31%	35%
Bacteroidota	1	5%	5%

**NA-Sequência não classificada na base de dados SILVA.*

Ao nível de género (**figura 17**), o grau de diferenciação aumenta bastante a nível das estirpes contaminantes. Foram identificados 35 géneros contaminantes nos controlos da extração. Os géneros contaminantes mais abundantes foram *Variovorax* (24 amostras), *Bosea* (23 amostras) e *Pseudomonas* (20 amostras), com uma variação de contaminação relativa entre 0,02% e 100% para algumas amostras, conforme indicado na **tabela XII**. Alguns géneros, como é o caso dos géneros *Comamonas, Elizabethkingia e Pasteurella*, apenas foram detectados, cada um, em uma única amostra e com uma taxa de contaminação de 0,1%, 5% e 0,3%, respetivamente. No total dos 35 géneros identificados nos controlos da extração, 17 géneros não contaminaram nenhuma amostra em estudo, estando unicamente presentes nos controlos (*Agromyces, Arthrobacter, Bifidobacterium, Chloroplast, Escherichia-Shigella, Fimbriimonadales, Novosphingobium, Paracoccus, Rapidithrix, Rhodobacter, Sphingomonas, Staphylococcus, Streptomyces, Unknown_Family, Vermiphilaceae, Vicinabacteraceae e Streptomyces).*

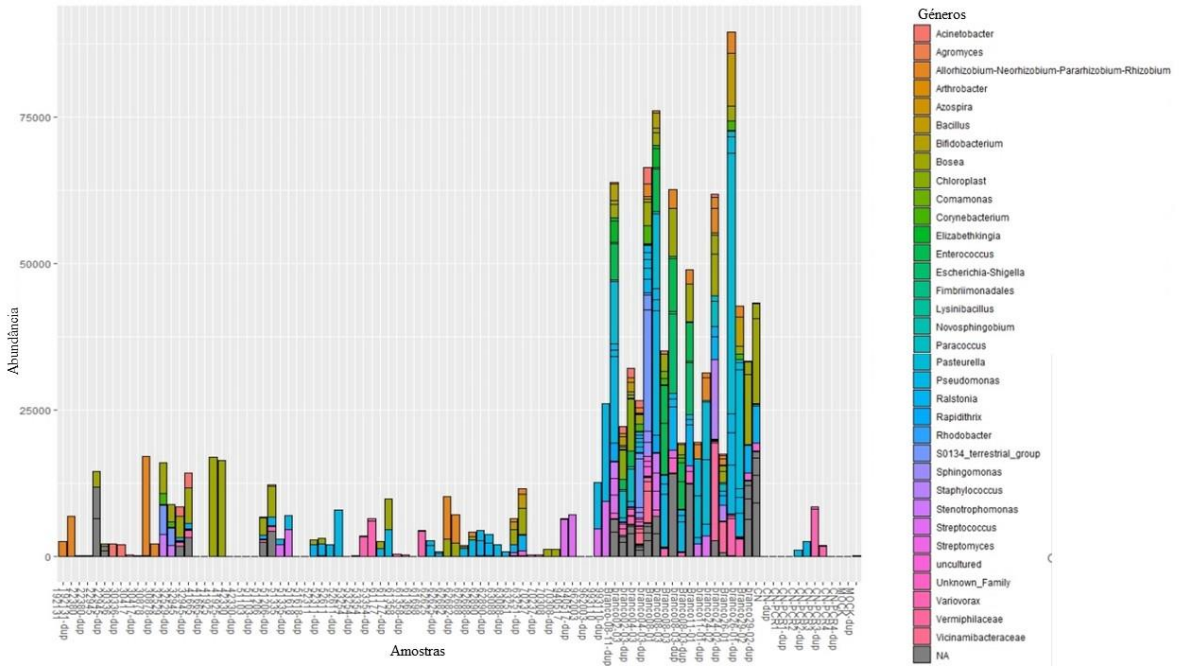


Figura 17. Abundância relativa dos géneros contaminantes presentes nos controlos da extração, nos 4 grupos em estudo

.

Tabela XII. Frequência e proporção relativa de géneros contaminantes identificados nos controlos da extração, que se encontravam presentes nas amostras.

Géneros Contaminantes	**Frequência**	**Mínimo**	**Máximo**
Variovorax	24	0.02%	100%
Bosea	23	14%	100%
Pseudomonas	20	0,2%	100%
Allorhizobium-Neorizobium-Pararhizobium-Rhizobium	12	0%	100%
Ralstonia	11	4%	73%
*NA**	8	0%	85%
Acinetobacter	4	0%	100%
Uncultured	4	7%	100%
Bacillus	3	20%	100%
Azospira	2	2%	2%
Corynebacterium	2	11%	12%
Enterococcus	2	0,5%	1%
Lysinibacillus	2	0,12%	0,14%
SO134_terrestrial_group	2	31%	35%
Stenotrophomonas	2	20%	23%
Comamonas	1	0,1%	0,1%
Elizabethkingia	1	5%	5%
Pasteurella	1	0,3%	0,3%

**NA-Sequência não classificada na base de dados SILVA.*

Nas **figuras 18** e **19** estão representadas graficamente as proporções relativas de cada filo e género contaminante, que foram identificados nos controlos da extracção para todas as 96 amostras deste estudo. Na figura 19 observa-se que o filo *Proteobacteria* apresenta uma grande proporção de contaminação na grande maioria das amostras em que foi identificado. Na figura 20 observa-se uma elevada diversidade de géneros contaminantes e, mesmo para amostras que tinham uma grande contaminação a nível do filo *Proteobacteria*, observam-se diferentes géneros contaminantes, isto é, existe uma relação de 1 filo – vários géneros.

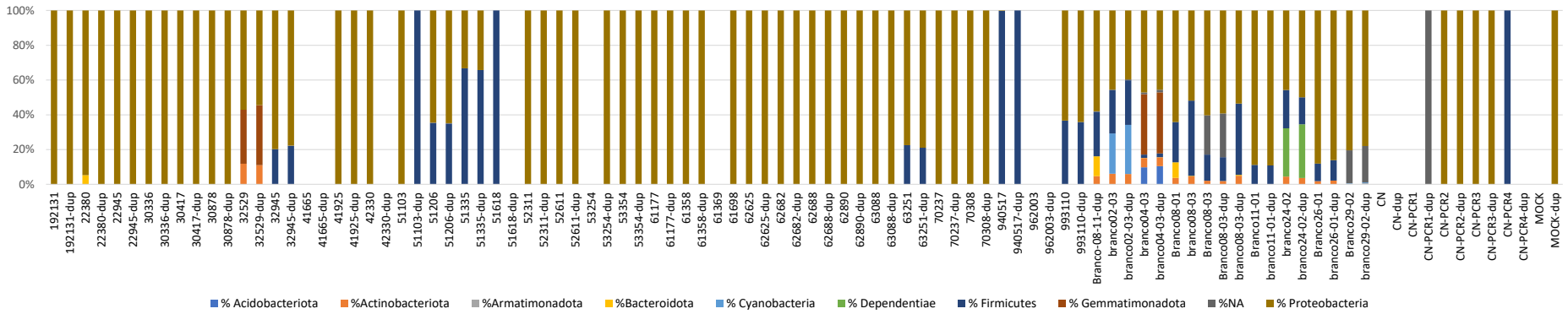


Figura 18. Representação gráfica da proporção relativa de filos contaminantes dos controlos da extração, nas amostras do estudo.

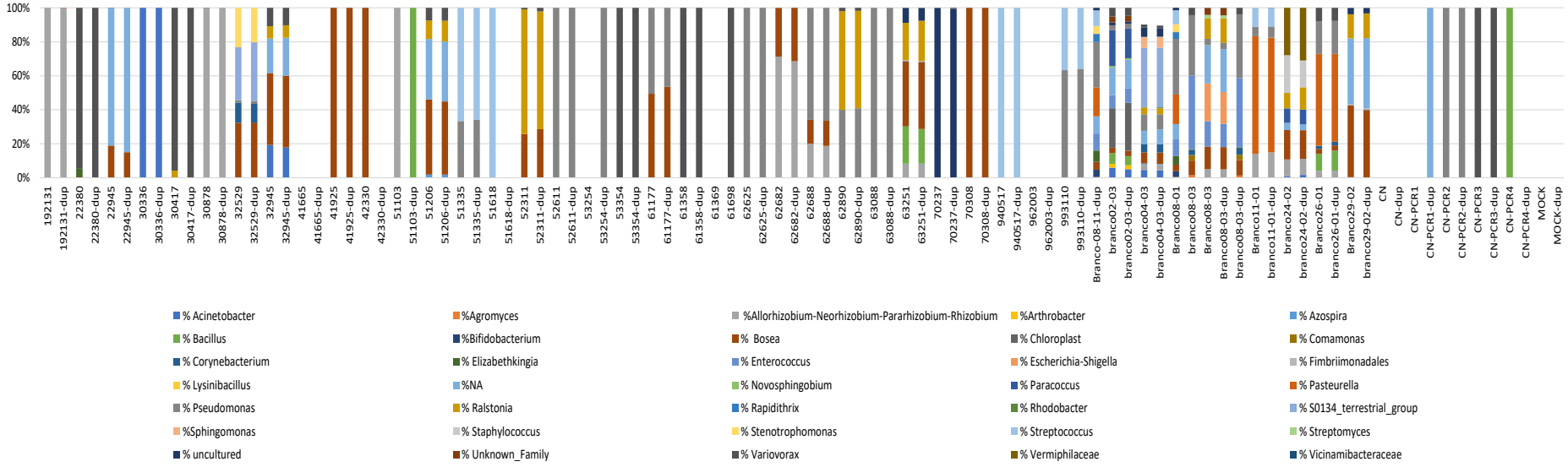


Figura 19. Representação gráfica da proporção relativa de géneros contaminantes dos controlos da extração, nas amostras do estudo.

5.4.2 Identificação dos contaminantes dos controlos da PCR

Foram também identificados e filtrados os contaminantes dos controlos da PCR, com o objetivo de identificar as estirpes contaminantes e caracterizar a sua abundância nas amostras do estudo. Ao nível de filo, obtiveram-se apenas 3 filos contaminantes, *Actinobacteriota, Firmicutes e Proteobacteria*, sendo as proteobactérias também as mais abundantes nos controlos negativos da PCR (**figura 20**).

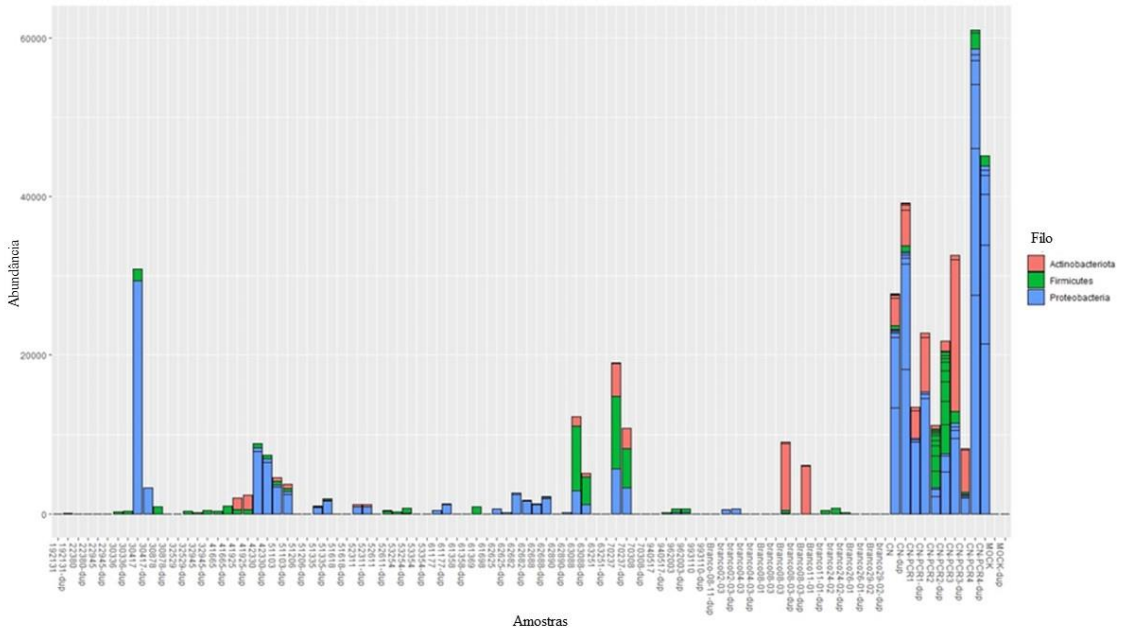


Figura 20. Abundância relativa de filos contaminantes dos controlos da PCR, presentes nas amostras em estudo.

Nos gráficos de frequências dos contaminantes presentes nos controlos negativos da PCR, verificou-se que no grupo das amostras, 22 destas não evidenciaram qualquer tipo de contaminação proveniente dos controlos negativos da PCR. Os filos mais dominantes foram as *proteobactérias*, que se encontraram presentes em 31 amostras, representando entre 3% e 100% de contaminação relativa, seguido pelo filo *Firmicutes*, que foi detectado em 30 amostras com uma representação relativa entre 0,4% e 100%. O filo *Actinobacteria* foi o contaminante menos frequente, estando apenas presente em 12 amostras e com uma contaminação relativa entre 12% e 79% (**tabela XIII**).

Tabela XIII. Frequência e proporção relativa de filos contaminantes identificados nos controlos da PCR, que se encontravam presentes nas amostras.

Filos Contaminantes	Frequência	Mínimo	Máximo
Proteobacteria	31	3%	100%
Firmicutes	30	0,4%	100%
Actinobacteria	12	12%	79%

Ao nível taxonómico de género, obteve-se maior diferenciação de estirpes, tendo sido identificados 17 géneros nos controlos negativos da PCR. Destes, os géneros mais abundantes foram *Variovorax* (20 amostras) e *Lysinibacillus* (19 amostras), com uma variação de contaminação entre 0,1%-100% e 5%-100% respectivamente, conforme indicado na **figura 21** e na **tabela XIV**.

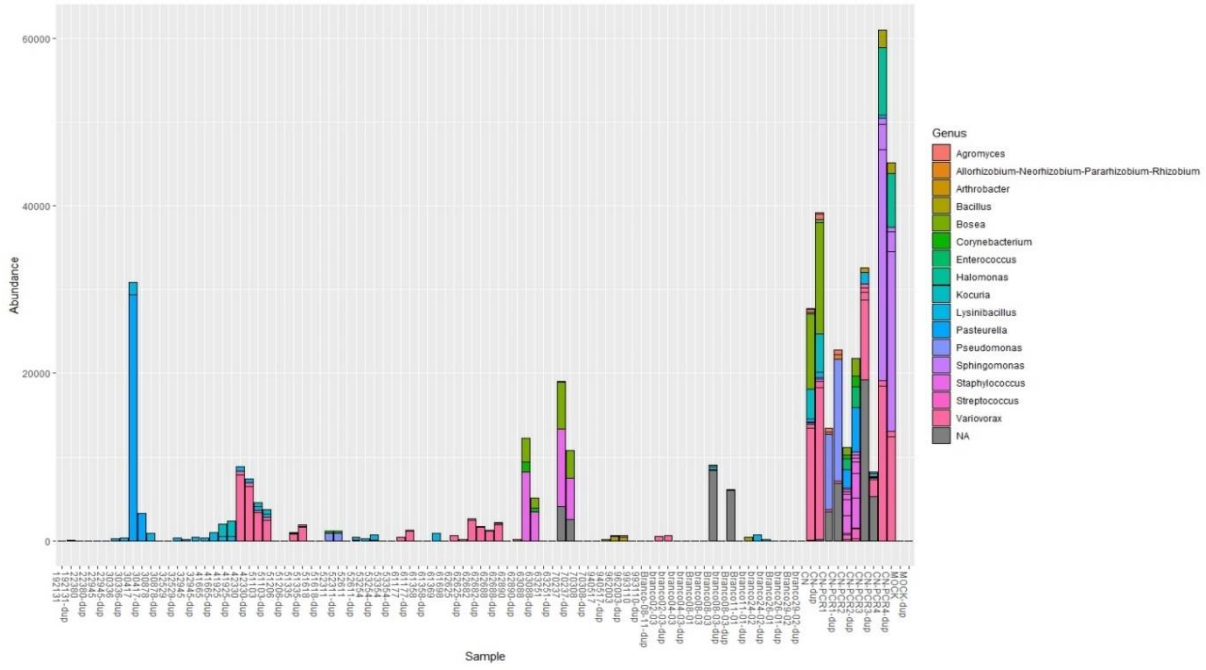


Figura 21. Abundância relativa de géneros contaminantes dos controlos da PCR, presentes nas amostras em estudo.

Os géneros *Agromyces*, *Enterococcus* e *Halomonas* não contaminaram nenhuma amostra do estudo e os restantes géneros identificados tiveram uma frequência de contaminação igual ou inferior a 5 amostras e com variação de contaminações entre 0,4% e 100% (**tabela XIV**).

Tabela XIV. Frequências e proporções relativas de géneros contaminantes identificados nos controlos da PCR, presentes nas amostras em estudo.

Géneros Contaminantes	**Frequência**	**Mínimo**	**Máximo**
Variovorax	20	0,1%	100%
Lysinibacillus	19	5%	100%
Staphylococcus	5	1%	68%
Bosea	4	23%	31%
Corynebacterium	4	9%	25%
Kocuria	4	11%	79%
*NA**	4	1%	24%
Pasteurella	4	3%	100%
Streptococcus	4	0,4%	10%
Allorhizobium-Neorizobium-Pararhizobium-Rhizobium	3	24%	100%
Bacillus	3	74%	94%
Pseudomonadaceae	2	74,6%	75%
Arthrobacter	1	1%	1%
Sphingomonas	1	6%	6%

**NA-Sequência não classificada na base de dados SILVA.*

Nas **figuras 22** e **23** são representadas as proporções relativas de cada filo e género contaminante, que foram identificados nos controlos da PCR para as 96 amostras do estudo. Na figura 22, com algumas excepções, observa-se que os filos *Proteobacteria* ou *Firmicutes*, representam a maioria, senão a totalidade dos contaminantes, presentes em cada amostra. Dado que o número de géneros contaminantes presentes nos controlos da PCR é menor que o encontrado nos controlos da extracção, as proporções relativas dos géneros contaminantes naqueles controlos parecem mimetizar melhor as proporções relativas dos respectivos filos, ou seja, há preferencialmente uma relação 1 filo – 1 género.

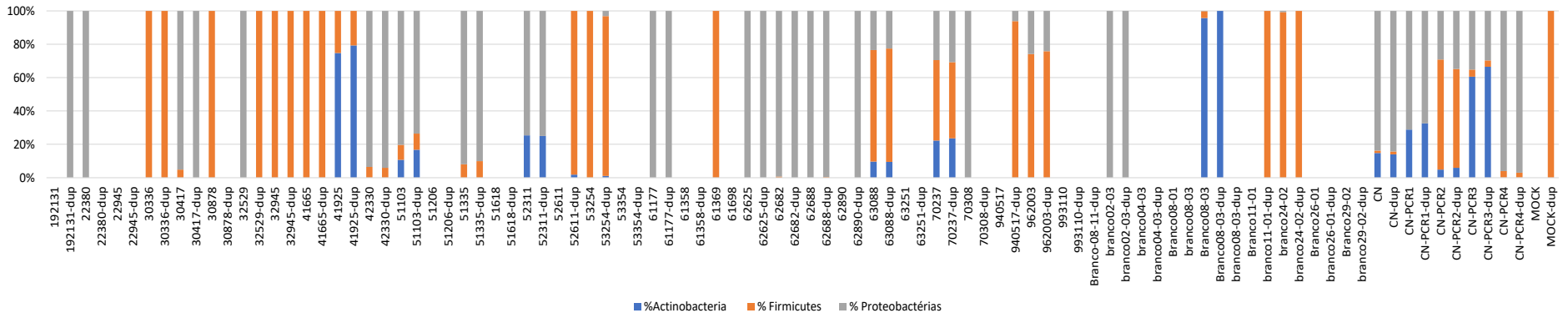


Figura 22. Representação gráfica da proporção relativa de filos contaminantes dos controlos da PCR, nas amostras do estudo.

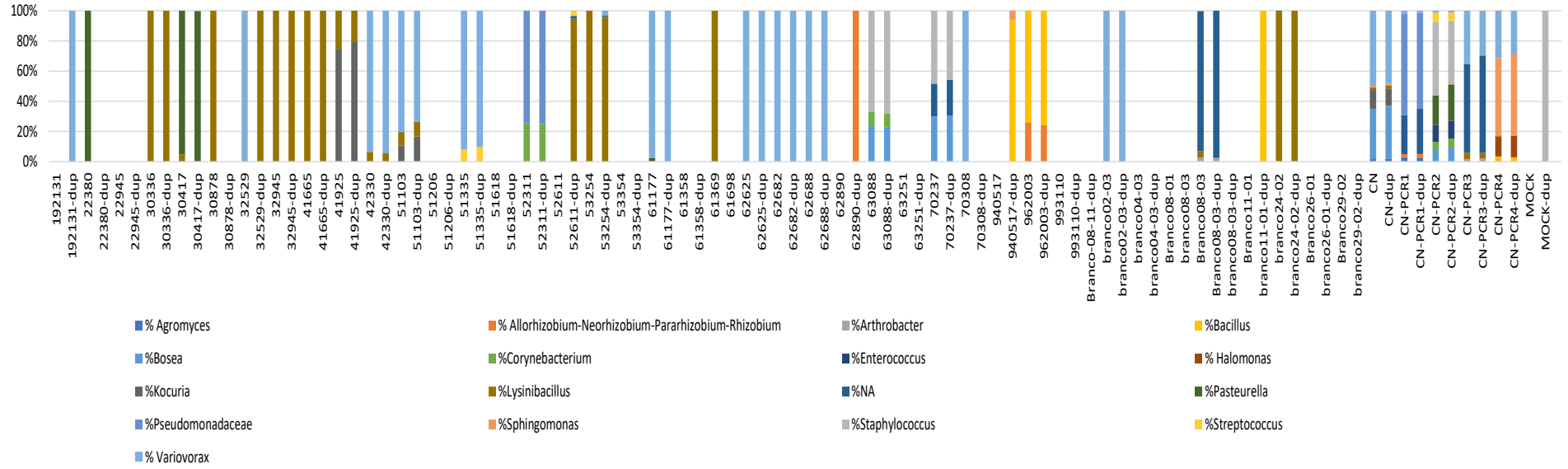


Figura 23. Representação gráfica da proporção relativa de géneros contaminantes dos controlos da PCR, nas amostras do estudo.

5.4.3 Impacto dos contaminantes na comunidade microbiana das amostras

Os taxa contaminantes foram identificados nas amostras com base na presença nos controlos negativos da extração de DNA e nos controlos negativos da PCR. Quando comparado o contributo destes contaminantes nas amostras, verificou-se que 4 conjuntos de amostras (30417/30417-dup, 51206/51206-dup, 70237/70237-dup e 993110/993110-dup) os contaminantes representaram aproximadamente 20-25% de contaminação face ao conteúdo taxonómico total da amostra, tendo este sido o maior valor de contaminação obtido nas amostras em estudo. As restantes amostras apresentaram níveis de contaminação abaixo dos 20%, com algumas amostras a ter níveis de contaminação pouco significativos em relação ao total do conteúdo taxonómico da amostra. Também se verificou que o principal momento de introdução de contaminação nas amostras aconteceu durante as extrações de DNA, tendo-se obtido uma maior percentagem de contaminação relativa em relação aos contaminantes introduzidos na etapa da PCR (**figura 24**). Este fato sugere assim que o momento da extração de DNA é mais suscetível de introdução de DNA bacteriano. Estas contaminações têm sido identificadas e relatadas também em diversos estudos de microbioma, com proveniência de fontes de contaminação que incluem reagentes de extrações de DNA, reagentes da PCR e outras contaminações ambientais, conforme tabela em anexo (**Anexo D**). No entanto, alguns taxa identificados neste estudo como contaminantes dos controlos negativos não foram identificados como contaminantes na bibliografia consultada, podendo constituir novos contaminantes específicos das condições ambientais/laboratoriais em que decorreu este estudo. Estes "novos" contaminantes incluem os géneros *Agromyces, Fimbriimonadales, Lysinibacillus, Rapidithrix, Rhodobacter, S134_terrestrial_group, Streptomyces, Vermiphilaceae e Vicinabacteraceae.* Além dos contaminantes presentes nas amostras, que foram identificados com base nos controlos da extracção e nos controlos da PCR, é legítimo admitir que as restantes estirpes presentes nas amostras de DNA tenham origem também em contaminações ambientais que ocorreram previamente à extracção de DNA. Dado que as amostras de tecido tumoral renal não foram colhidas e armazenadas em condições controladas para estudos de microbioma, e que a presença de microorganismos simbiontes das células renais humanas não é expectável, ou é residual, a presença de outros contaminantes pré-extracção de DNA é uma forte possibilidade. No entanto, por não fazer parte do âmbito deste trabalho, a caracterização dessas estirpes deverá constituir um objecto de investigação futura.

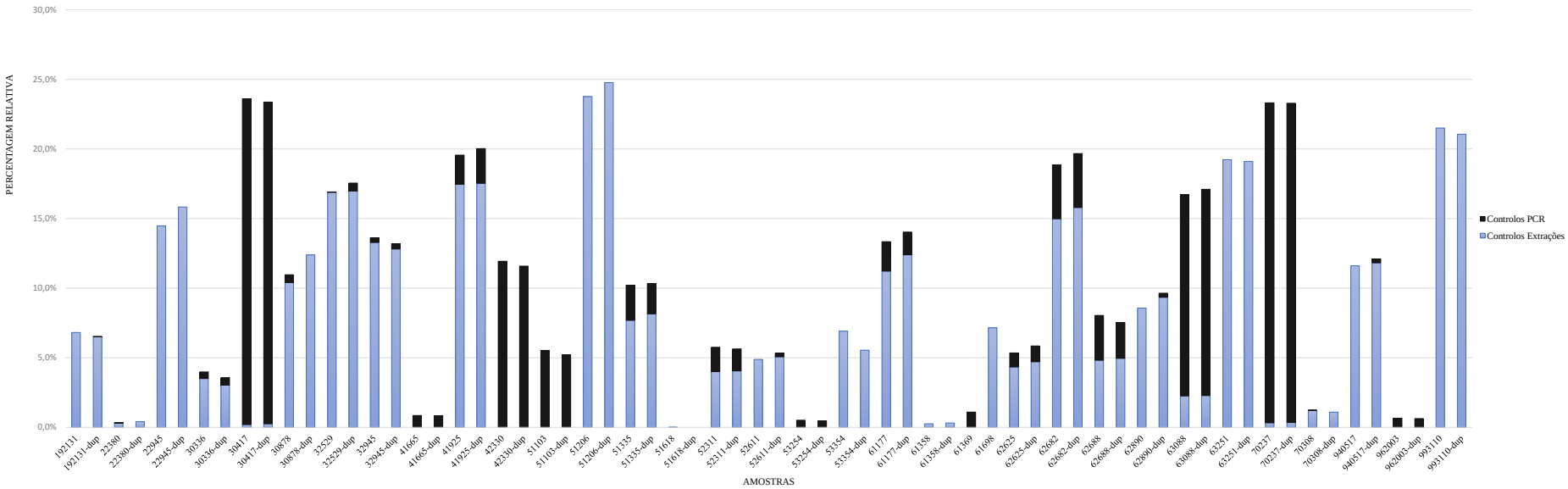


Figura 24. Representação gráfica do contributo relativo dos contaminantes das extrações e dos contaminantes da PCR nas amostras.

6 Análise do controlo mock

Da análise taxonómica do controlo "mock" foi possível identificar ao nível de espécie apenas a bactéria *Listeria monocytogenes*. As restantes 3 estirpes apenas foram identificadas ao nível de ordem para *Mycobacterium tuberculosis* e *Neisseria gonorrhoeae*, e ao nível de género para *Streptococus pneumoniae.* O facto de o QIIME2 não conseguir atribuir correctamente a taxonomia ao nível de espécie pode acontecer devido a duas situações, relativas à sequenciação das regiões V3 e V4 do gene 16S rRNA não ter poder discriminatório suficiente ou a possíveis erros de sequenciação que dificultaram a atribuição da classificação taxonómica específica com a base de dados utilizada.

Também as proporções relativas das 4 estirpes na amostra "mock" foram desiguais, ainda que se tenham adicionado na mesma concentração/volume. Estas diferenças de proporções podem acontecer na medida em que, ainda que se adicione a mesma quantidade, equilibramos a sua massa, mas não o número de cópias do genoma, uma vez que se tratam de bactérias com diferentes tamanhos de genoma. Este facto gerou assim um desequilíbrio das suas proporções. Neste sentido, uma comunidade "mock" deve ser construída de acordo com o tamanho dos genomas usados.

7 Discussão

A vantagem de efetuar estudos metagenómicos em DNA obtido diretamente da extração de amostras, sem necessidade de crescimento de culturas, veio simplificar os procedimentos laboratoriais e permitir a identificação de um grande número de microrganismos não-cultiváveis. A possibilidade de sequenciar uma região alvo, com *primers* específicos para as regiões hipervariáveis V3 e V4 do gene 16S rRNA, permitiu fornecer descrições detalhadas de comunidades microbianas nas amostras em estudo, e vários trabalhos referem que estas regiões são mais representativas para estudos metagenómicos [27]. No entanto, os estudos metagenómicos, em consequência da utilização de amplificação de DNA na preparação de amostras para sequenciação, e do elevado rendimento das plataformas de sequenciação paralela massiva, tem levado a problemas de contaminações ambientais possíveis de serem introduzidas nas amostras em estudo.

No presente trabalho a análise bioinformática das sequências obtidas do gene 16S rRNA gerou grandes desafios até obter a classificação taxonómica e poder inferir quais as contaminações introduzidas durante o processamento das amostras. Esta análise compreendeu a utilização de dois *softwares:* o *QIIME2* para obtenção da taxonomia, e o Decontam para determinação e remoção de taxa contaminantes. Ambos os programas utilizados são de acesso gratuito e de fácil compreensão, no entanto foi necessário despender algum tempo para a sua realização e otimização de acordo com as amostras do estudo.

Ao nível do *QIIME2*, as primeiras etapas da análise foram efetuadas com o *plugin* DADA2, com o objetivo de remover e corrigir erros de sequenciação, e filtrar sequencias quiméricas e *reads* de baixa qualidade, que não afetassem posteriormente a atribuição da taxonomia às amostras. Este *plugin* foi testado com diferentes valores para os parâmetros de truncagem no final das *reads* (de acordo com os gráficos de qualidade da corrida), remoção de sequências de baixa qualidade, *trimming* no início das *reads*, e número máximo de erros permitidos, de forma a entender quais os valores que resultavam no maior número final de *reads*. Os vários testes demonstraram que truncagens mais conservadoras causavam maior impacto nos valores de filtragem, retendo a maior parte das *reads* e, consequentemente, obtendo menos *reads* no final. A variação do número de erros permitidos não teve efeito relevante no número final de *reads*, no entanto o *trimming* das *reads* teve um impacto significativo na redução do número de sequências identificadas como quiméricas. Após tratamento das *reads*, ainda se manteve uma quantidade de *reads* elevada (~6.8M) que permitiu reunir de forma sensível e robusta a informação sobre o número de vezes que cada sequencia (ASV) foi observada em cada amostra, para prosseguir com as análises seguintes.

Para responder às questões sobre quantas espécies/ASV diferentes estariam presentes no estudo, e qual a diversidade presente nas amostras, foram realizadas análises de diversidade. Das análises de diversidade alfa, foram observados os índices de *Faith* (riqueza) e de *Shannon* (diversidade), demonstrando que as amostras possuíam maior riqueza de ASVs/espécies e diversidade em comparação com os controlos usados (extração, PCR e controlo positivo), e que os controlos da extração também já sugeriam ter maior diversidade em relação aos controlos da PCR. A rarefação alfa efetuada também corroborou que a profundidade de sequência escolhida, demonstrava representatividade e eficácia na captura de toda a diversidade presente nos controlos. As análises de diversidade beta também indicaram algumas diferenças na composição de espécies entre as amostras, nomeadamente a distância de *Jaccard,* quando comparadas as distâncias dos diversos controlos em relação às amostras. Para este índice, foi demonstrado que, quer os controlos da extração, da PCR ou até o controlo positivo compartilhavam poucas espécies em relação à maioria das amostras.

Quando atribuída a taxonomia às amostras do estudo, esta corroborou as informações extraídas das análises de diversidade alfa, nomeadamente que as amostras em estudo apresentavam uma grande riqueza ao nível de espécie. Observou-se também que ao nível de filo as *Proteobacteria* estavam presentes em todas as amostras e controlos incluídos no estudo. No entanto, a visualização dos taxa por grupos tornou-se mais difícil, não permitindo perceber quais os taxa contaminantes de cada controlo que estavam presentes em cada uma das amostras. Para a identificação e remoção de taxa contaminantes utilizou-se o Decontam um pacote que trabalha em R, com uma interface simples e que fornece métodos estatísticos que permitem a remoção de contaminantes, melhorando a visualização de dados metagenómicos. A utilização do Decontam, permitiu identificar os taxa contaminantes presentes nas amostras dos controlos negativos (controlo da extração e controlo da PCR). Como era de esperar pelas análises de diversidade, os controlos negativos das extrações apresentaram contaminações com maior diversidade taxonómica em relação aos controlos negativos da PCR, assim como as contaminações identificadas em ambos os controlos afetaram desproporcionalmente as amostras do estudo. Ao nível de filo, entre os 10 filos identificados como contaminantes das extrações, apenas um deles (estando apenas presente em 3 amostras e com uma variação de contaminação entre 0,2% e 0,3%) não foi classificado na base de dados SILVA. As *Proteobacteria* foram o filo mais abundante e frequente, estando presente em 55 amostras do estudo (no total de 66 amostras) e representando na maior parte delas 100% de contaminação, ou seja, este era o único filo contaminante das amostras. De todos os filos identificados como contaminantes, as *Acidobacteriota, Armatimonadota, Cyanobacteria e Dependentiae* não apresentaram sequências contaminantes nas amostras, embora representassem 40% de filos contaminantes exclusivos dos controlos negativos das extrações. Ao nível do género, 48% dos géneros classificados como contaminantes não foram observados nas amostras. Os géneros mais frequentes nas amostras foram *Variovorax* (n=24), *Bosea* (n=23), *Pseudomonas* (n=20), *Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium* (n=12) *e Ralstonia* (n=11), tendo os restantes géneros uma frequência entre 1-8 amostras.

Nos controlos negativos da PCR, obtiveram-se menos contaminantes em relação aos controlos negativos das extrações. Ao nível de filo, obtiveram-se apenas 3 filos contaminantes entre as amostras, sendo novamente as *Proteobacteria* as mais abundantes e frequentes (n=31), seguido de *Firmicutes* (n=30) e, com menor abundância/frequência, *Actinobacteria* (n=12). Apesar de terem sido identificados 17 géneros contaminantes, 3 não demonstraram qualquer contaminação nas amostras (*Agromyces, Enterococcus e Halomonas)* e apenas 2 tiveram uma maior frequência e abundância, *Variovorax* (n=20) e *Lysinibacillus* (n=19).

Quando investigado qual o contributo dos contaminantes na comunidade microbiana das amostras, este revela-se no geral inferior a 25% de contaminação e observa-se que o maior impacto na introdução de contaminantes acontece durante as extracções de DNA. Uma justificação para o facto de se terem obtido menos contaminantes nos controlos da PCR em relação aos controlos negativos da extração, pode ser devido à separação de áreas laboratoriais, uma vez que a preparação das misturas da PCR é efetuada em laboratório com áreas limpas e onde não são manuseados DNA´s. Alguns autores atribuem a designação de "kitoma" (contaminação introduzida por kits/reagentes de extração de DNA) às contaminações detetadas nos controlos da extração [28]. No entanto, alguns géneros identificados como contaminantes presentes nos controlos negativos (como o caso dos géneros A*gromyces, Fimbriimonadales, Lysinibacillus, Rapidithrix, Rhodobacter, S134_terrestrial_group, Streptomyces, Vermiphilaceae e Vicinabacteraceae),* não foram identificados como contaminantes em estudos de microbioma, nas pesquisas bibliográficas efetuadas. A maioria destes contaminantes aparecem descritos como isolados de amostras de solo e de água doce, podendo constituir novos contaminantes presentes no ambiente. Curiosamente dois géneros contaminantes, *Rapidithrix* e *Streptomyces* são descritas como bactérias usadas em produção de antibióticos.

Além dos contaminantes presentes nas amostras, que foram identificados com base nos controlos da extracção e nos controlos da PCR, é legítimo admitir que as restantes estirpes presentes nas amostras de DNA tenham origem também em contaminações ambientais que ocorreram previamente à extracção de DNA. Dado que as amostras de tecido tumoral renal não foram colhidas e armazenadas em condições controladas para estudos de microbioma, e que a presença de microorganismos simbiontes das células renais humanas não é expectável, ou é residual, a presença de outros contaminantes pré-extracção de DNA é uma forte possibilidade. No entanto, por não fazer parte do âmbito deste trabalho, a caracterização dessas estirpes deverá constituir um objecto de investigação futura.

O Decontam demonstrou ser uma ferramenta eficaz para identificar e remover as sequencias contaminantes. No entanto, uma limitação que os autores referem e que se verificou no presente trabalho é que, ao assumir como verdadeiros contaminantes todas as sequências presentes nos controlos negativos, distinguindo-os das amostras, não tem em conta possíveis contaminações cruzadas decorrentes do processamento das amostras [15]. Um exemplo neste contexto foi o de se terem identificado taxa contaminantes em amostras cujo controlo negativo, utilizado na mesma data de extracção não possuía esses contaminantes, mas sim taxa de outra(s) data(s) de extração. Esta evidência leva a ter em atenção possíveis contaminações cruzadas ou até contaminações anteriores das amostras, uma vez que estas, para além da sua antiguidade, não foram colhidas e transportadas tendo em conta os objectivos deste estudo. Neste sentido, o objetivo foi cumprido, uma vez que consistiu em implementar uma metodologia de análise de microbioma e poder excluir possíveis contaminações introduzidas no processamento laboratorial. No entanto, e de acordo com vários estudos, para além da introdução de controlos negativos, outras recomendações devem ser levadas em conta para mitigar a introdução de DNA contaminante e contaminações cruzadas em amostras com baixa biomassa [29]. Para tal, devem ser tomadas medidas restritivas desde a colheita da amostra até ao seu processamento final, tais como:

- A colheita deve ser controlada e efetuada em ambiente limpo;
- Usar equipamento de proteção individual que cubra as áreas do corpo expostas;
- Limpar bancadas e superfícies com hipoclorito de sódio a 3% e radiação UV para diminuir contaminação ambiental;
- Utilizar consumíveis que sejam livres de DNA exógeno;
- Efetuar alíquotas de reagentes para minimizar manipulações desnecessárias e consequentes contaminações;
- Separar áreas de trabalho pré e pós PCR;
- Utilizar pontas com filtro e pipetas com baixo teor de aerossol;
- Fazer lavagens de manutenção do sequenciador com NaOCl (recomendação Illumina, para diminuição de contaminação cruzada com outras corridas de sequenciação);
- Introduzir controlos negativos por amostra;
- Utilizar, se possível, um controlo positivo comercial para permitir uma padronização entre laboratórios;

Todos estes procedimentos devem ser considerados para estudos com amostras de muito baixa biomassa microbiana. Além disso é necessário adotar uma estratégia balanceada de forma a obter-se uma maior sensibilidade de deteção de microrganismos, sem conduzir a uma maior contaminação ambiental das amostras. Em resumo, conclui-se que as estirpes contaminantes são um problema sério em estudos de microbioma humano com reduzida biomassa, e que a metodologia apresentada aqui é uma abordagem eficiente para detectar e quantificar essas estirpes.

8 Bibliografia

1. Grice, E. A. & Segre, J. A. The Human Microbiome: Our Second Genome. *Annu. Rev. Genomics Hum. Genet.* **13**, 151–170 (2012).
2. Ursell, L. K., Metcalf, J. L., Parfrey, L. W. & Knight, R. Defining the Human Microbiome. *Nutr Rev* **70**, s38–s44 (2012).
3. Martín, R., Miquel, S., Langella, P. & Bermúdez-Humarán, L. G. The role of metagenomics in understanding the human microbiome in health and disease. *Virulence* **5**, 413–423 (2014).
4. Funkhouser, L. J. & Bordenstein, S. R. Mom Knows Best: The Universality of Maternal Microbial Transmission. *PLoS Biol.* **11**, 1–9 (2013).
5. Lim, E. S., Rodriguez, C. & Holtz, L. R. Amniotic fluid from healthy term pregnancies does not harbor a detectable microbial community. *Microbiome* **7**, 4–11 (2019).
6. Parfrey, L. W. & Knight, R. Spatial and temporal variability of the human microbiota. *Clin. Microbiol. Infect.* **18**, 8–11 (2012).
7. Turnbaugh, P. J. *et al.* The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* **449**, 804–810 (2007).
8. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Narure Biotechnology* vol. 26 1135–1145 (2008).
9. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem. Biol.* **5**, (1998).
10. Hugenholtz, P., Goebel, B. M. & Pace, N. R. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *J. Bacteriol.* **180**, 6793 (1998).
11. Woese, C. R. *et al.* Conservation of primary structure in 16S ribosomal RNA. *Nature* **254**, 83–86 (1975).
12. Větrovský, T. & Baldrian, P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS One* **8**, 1–10 (2013).
13. J Gregory, C. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**, 335–336 (2010).
14. Callahan, B. J. *et al.* Dada2 High resolution sample inference from Illumina amplicon data. *Nat Methods.* **13**, 581–583 (2016).
15. Davis, N. M., Proctor, Di. M., Holmes, S. P., Relman, D. A. & Callahan, B. J. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 1–14 (2018).
16. Karstens, L. *et al.* Controlling for Contaminants in Low-Biomass 16S rRNA Gene Sequencing Experiments. *mSystems* **4**, 1–14 (2019).
17. Philip Ewels, Måns Magnusson, S. L. and M. K. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatic* (2016).

18. Andrews, S. FastQC: A quality control tool for high throughput sequence data. *http//www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed 2 August 2017).* (2010).

19. Faith, D. P. Conservation evaluation and phylogenetic diversity. *Biol. Conserv.* **61**, 1–10 (1992).

20. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41**, 590–596 (2013).

21. Navas-Molina, J. A. *et al.* Advancing our understanding of the human microbiome using QIIME. *Methods Enzym.* 371–444 (2013) doi:10.1016/B978-0-12-407863-5.00019-8.

22. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27**, 623–656 (1948).

23. Spellerberg, I. F. & Fedor, P. J. A tribute to Claude-Shannon (1916-2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon-Wiener' Index. *Glob. Ecol. Biogeogr.* **12**, 177–179 (2003).

24. Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. (1901) doi:10.5169/seals-266450.

25. Lozupone, C. A. & Knight, R. Global patterns in bacterial diversity. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 11436–11440 (2007).

26. Weyrich, L. S. *et al.* Laboratory contamination over time during low-biomass sample analysis. *Mol. Ecol. Resour.* **19**, 982–996 (2019).

27. Walters, W. *et al.* Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. *mSystems* **1**, 1–10 (2016).

28. Stinson, L. F., Keelan, J. A. & Payne, M. S. Identification and removal of contaminating microbial DNA from PCR reagents: impact on low-biomass microbiome analyses. *Lett. Appl. Microbiol.* **68**, 2–8 (2019).

29. Goodrich, J. K. *et al.* Conducting a microbiome study. *Cell* **158**, 250–262 (2014).

30. Tanner, M. A., Goebel, B. M., Dojka, M. A. & Pace, N. R. Specific ribosomal DNA sequences from diverse environmental settings correlate with experimental contaminants. *Appl. Environ. Microbiol.* **64**, 3110–3113 (1998).

31. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12**, 1–12 (2014).

32. Lauder, A. P. *et al.* Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome* **4**, 1–11 (2016).

33. Saladié, M. *et al.* Microbiomic Analysis on Low Abundant Respiratory Biomass Samples; Improved Recovery of Microbial DNA From Bronchoalveolar Lavage Fluid. *Front. Microbiol.* **11**, 1–11 (2020).

34. Glassing, A., Dowd, S. E., Galandiuk, S., Davis, B. & Chiodini, R. J. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog.* **8**, 1–12 (2016).

ANEXOS

Anexo A – Gráficos representativos do MultiQC report

Instituto_Nacional de Saúde
Doutor Ricardo Jorge

MultiQC

(http://multiqc.info)

Análise da qualidade da sequenciação

Dep. de Genética Humana - Unidade de Tecnologia e Inovação

A análise primária foi efectuada usando os programas Interop e FastQC

Report generated on 2019-07-12, 13:10 based on data in: `/mnt/san/qc_tmp_files/190709_M01600_0112_000000000-CCC94`

Sequence Quality Histograms 0 192

The mean quality value across each base position in the read. See the FastQC help (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/2%20Per%20Base%20Sequence%20Quality.html).

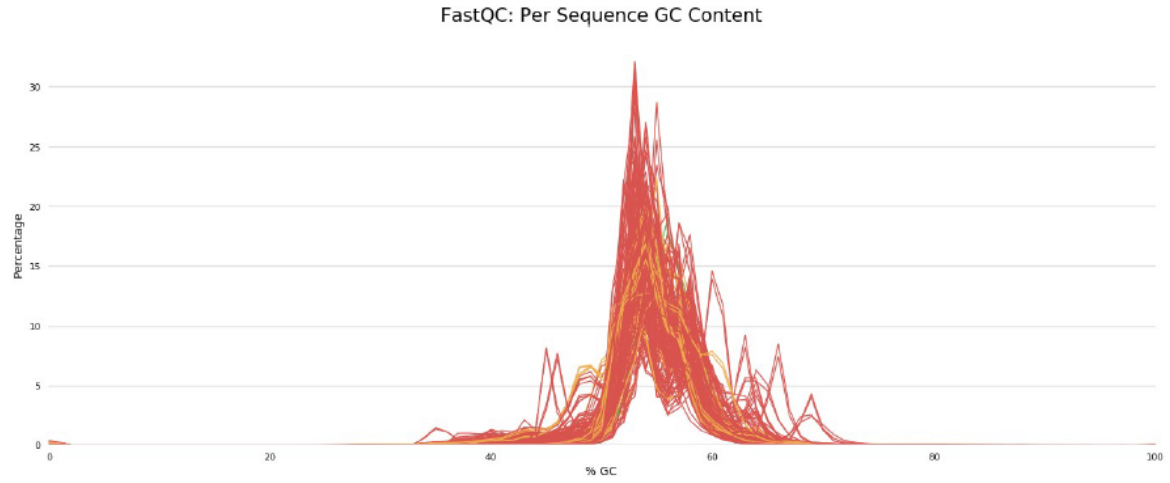
Flat image plot. Toolbox functions such as highlighting / hiding samples will not work (see the docs (http://multiqc.info/docs/#flat--interactive-plots)).

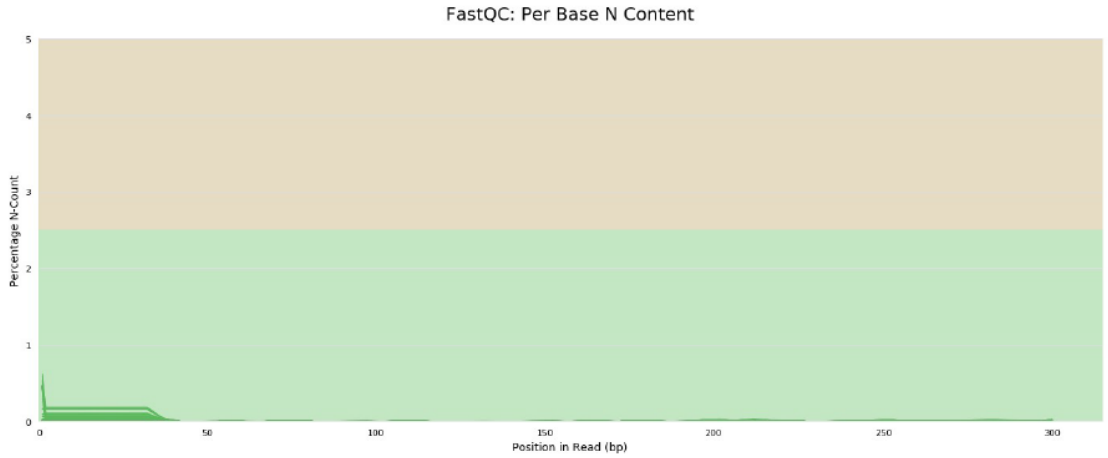
Per Sequence GC Content 32 158

The average GC content of reads. Normal random library typically have a roughly normal distribution of GC content. See the FastQC help (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/5%20Per%20Sequence%20GC%20Content.html).

Flat image plot. Toolbox functions such as highlighting / hiding samples will not work (see the docs (http://multiqc.info/docs/#flat--interactive-plots)).

Percentages | Counts

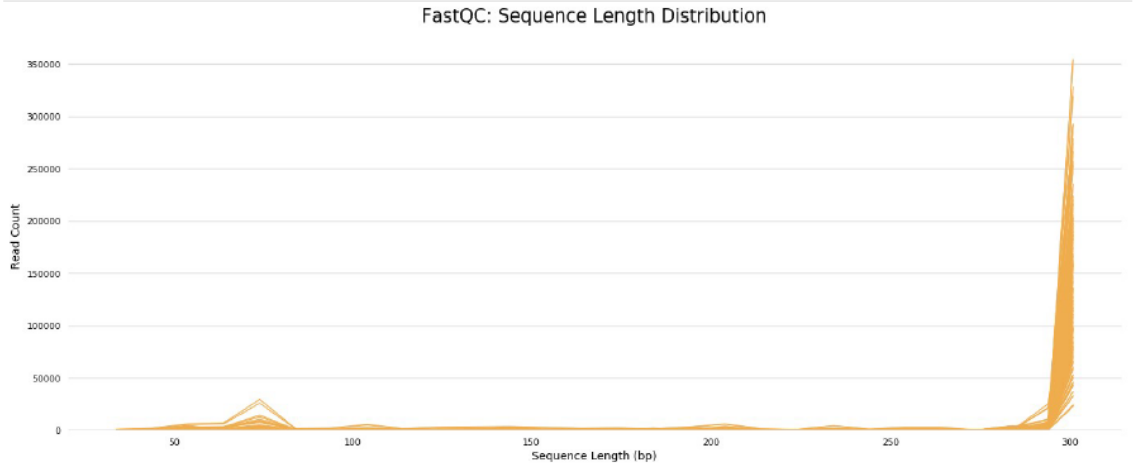




Sequence Length Distribution 0 192

The distribution of fragment sizes (read lengths) found. See the FastQC help (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/7%20Sequence%20Length%20Distribution.html).

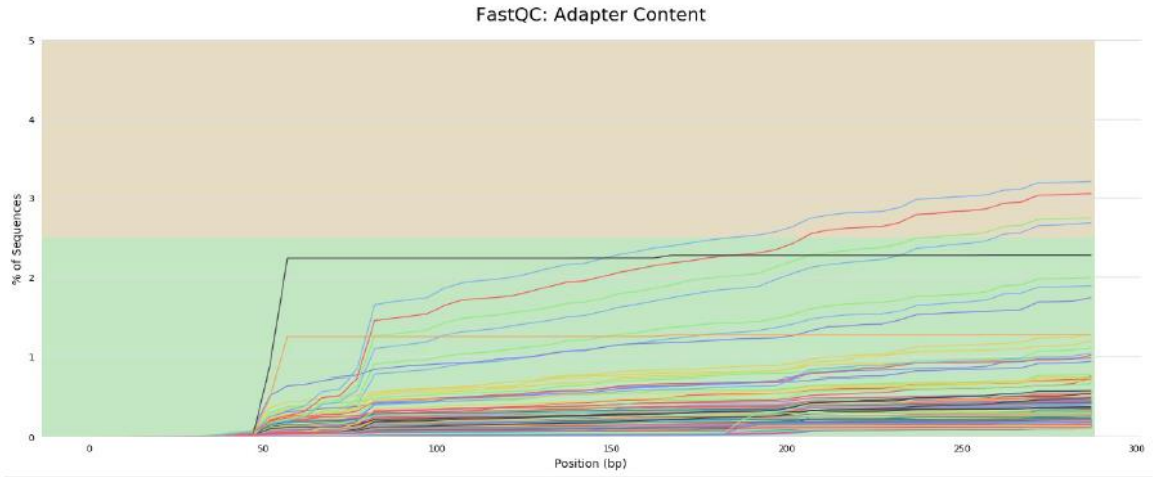
Flat image plot. Toolbox functions such as highlighting / hiding samples will not work (see the docs (http://multiqc.info/docs/#flat--interactive-plots)).



Adapter Content 192

The cumulative percentage count of the proportion of your library which has seen each of the adapter sequences at each position. See the FastQC help (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/10%20Adapter%20Content.html). Only samples with ≥ 0.1% adapter contamination are shown.

Flat image plot. Toolbox functions such as highlighting / hiding samples will not work (see the docs (http://multiqc.info/docs/#flat--interactive-plots)).



Anexo B – Ficheiro de metadados com informação das amostras em estudo

sample-id	I7_Index_ID	index	I5_Index_ID	index2	year	month	day	replicate	Sample_Type
22380	N704	TCCTGAGC	S508	CTAAGCCT	2019	1	29	1	sample
22945	N705	GGACTCCT	S503	TATCCTCT	2019	2	1	1	sample
30336	N705	GGACTCCT	S504	AGAGTAGA	2019	2	1	1	sample
30417	N705	GGACTCCT	S505	GTAAGGAG	2019	2	1	1	sample
30878	N705	GGACTCCT	S506	ACTGCATA	2019	2	1	1	sample
32529	N703	AGGCAGAA	S508	CTAAGCCT	2019	1	29	1	sample
32945	N704	TCCTGAGC	S517	GCGTAAGA	2019	1	29	1	sample
41665	N704	TCCTGAGC	S503	TATCCTCT	2019	1	29	1	sample
41925	N704	TCCTGAGC	S502	CTCTCTAT	2019	1	29	1	sample
42330	N704	TCCTGAGC	S505	GTAAGGAG	2019	1	29	1	sample
51103	N704	TCCTGAGC	S506	ACTGCATA	2019	1	29	1	sample
51206	N704	TCCTGAGC	S507	AAGGAGTA	2019	1	29	1	sample
51335	N702	CGTACTAG	S517	GCGTAAGA	2019	1	24	1	sample
51618	N702	CGTACTAG	S502	CTCTCTAT	2019	1	24	1	sample
52311	N702	CGTACTAG	S503	TATCCTCT	2019	1	24	1	sample
52611	N702	CGTACTAG	S504	AGAGTAGA	2019	1	24	1	sample
53254	N706	TAGGCATG	S517	GCGTAAGA	2019	2	1	1	sample
53354	N706	TAGGCATG	S502	CTCTCTAT	2019	2	1	1	sample
61177	N703	AGGCAGAA	S505	GTAAGGAG	2019	1	24	1	sample
61358	N706	TAGGCATG	S503	TATCCTCT	2019	2	1	1	sample
61369	N706	TAGGCATG	S504	AGAGTAGA	2019	2	1	1	sample
61698	N706	TAGGCATG	S505	GTAAGGAG	2019	2	1	1	sample
62625	N702	CGTACTAG	S507	AAGGAGTA	2019	1	24	1	sample
62682	N702	CGTACTAG	S508	CTAAGCCT	2019	1	24	1	sample
62688	N703	AGGCAGAA	S517	GCGTAAGA	2019	1	24	1	sample
62890	N703	AGGCAGAA	S502	CTCTCTAT	2019	1	24	1	sample
63088	N703	AGGCAGAA	S503	TATCCTCT	2019	1	24	1	sample
63251	N703	AGGCAGAA	S504	AGAGTAGA	2019	1	24	1	sample
70237	N702	CGTACTAG	S505	GTAAGGAG	2019	1	24	1	sample
70308	N705	GGACTCCT	S508	CTAAGCCT	2019	2	1	1	sample
192131	N701	TAAGGCGA	S506	ACTGCATA	2019	1	18	1	sample
940517	N701	TAAGGCGA	S505	GTAAGGAG	2019	1	18	1	sample
962003	N701	TAAGGCGA	S503	TATCCTCT	2019	1	18	1	sample
993110	N701	TAAGGCGA	S517	GCGTAAGA	2019	1	18	1	sample
192131-dup	N707	CTCTCTAC	S507	AAGGAGTA	2019	1	18	2	sample
22380-dup	N711	AAGAGGCA	S517	GCGTAAGA	2019	1	29	2	sample
22945-dup	N711	AAGAGGCA	S504	AGAGTAGA	2019	2	1	2	sample
30336-dup	N711	AAGAGGCA	S505	GTAAGGAG	2019	2	1	2	sample
30417-dup	N711	AAGAGGCA	S506	ACTGCATA	2019	2	1	2	sample
30878-dup	N711	AAGAGGCA	S507	AAGGAGTA	2019	2	1	2	sample
32529-dup	N710	CGAGGCTG	S517	GCGTAAGA	2019	1	29	2	sample
32945-dup	N710	CGAGGCTG	S502	CTCTCTAT	2019	1	29	2	sample
41665-dup	N710	CGAGGCTG	S504	AGAGTAGA	2019	1	29	2	sample
41925-dup	N710	CGAGGCTG	S503	TATCCTCT	2019	1	29	2	sample
42330-dup	N710	CGAGGCTG	S506	ACTGCATA	2019	1	29	2	sample
51103-dup	N710	CGAGGCTG	S507	AAGGAGTA	2019	1	29	2	sample
51206-dup	N710	CGAGGCTG	S508	CTAAGCCT	2019	1	29	2	sample
51335-dup	N708	CAGAGAGG	S502	CTCTCTAT	2019	1	24	2	sample
51618-dup	N708	CAGAGAGG	S503	TATCCTCT	2019	1	24	2	sample
52311-dup	N708	CAGAGAGG	S504	AGAGTAGA	2019	1	24	2	sample
52611-dup	N708	CAGAGAGG	S505	GTAAGGAG	2019	1	24	2	sample
53254-dup	N712	GTAGAGGA	S502	CTCTCTAT	2019	2	1	2	sample
53354-dup	N712	GTAGAGGA	S503	TATCCTCT	2019	2	1	2	sample
61177-dup	N709	GCTACGCT	S506	ACTGCATA	2019	1	24	2	sample
61358-dup	N712	GTAGAGGA	S504	AGAGTAGA	2019	2	1	2	sample
62625-dup	N708	CAGAGAGG	S508	CTAAGCCT	2019	1	24	2	sample
62682-dup	N709	GCTACGCT	S517	GCGTAAGA	2019	1	24	2	sample
62688-dup	N709	GCTACGCT	S502	CTCTCTAT	2019	1	24	2	sample
62890-dup	N709	GCTACGCT	S503	TATCCTCT	2019	1	24	2	sample
63088-dup	N709	GCTACGCT	S504	AGAGTAGA	2019	1	24	2	sample
63251-dup	N709	GCTACGCT	S505	GTAAGGAG	2019	1	24	2	sample
70237-dup	N708	CAGAGAGG	S506	ACTGCATA	2019	1	24	2	sample
70308-dup	N712	GTAGAGGA	S517	GCGTAAGA	2019	2	1	2	sample
940517-dup	N707	CTCTCTAC	S506	ACTGCATA	2019	1	18	2	sample
962003-dup	N707	CTCTCTAC	S504	AGAGTAGA	2019	1	18	2	sample
993110-dup	N707	CTCTCTAC	S502	CTCTCTAT	2019	1	18	2	sample
branco02-03	N705	GGACTCCT	S517	GCGTAAGA	2019	1	29	1	Extraction_Control
branco02-03-dup	N711	AAGAGGCA	S502	CTCTCTAT	2019	1	29	2	Extraction_Control
branco04-03	N702	CGTACTAG	S506	ACTGCATA	2019	1	24	1	Extraction_Control
branco04-03-dup	N708	CAGAGAGG	S507	AAGGAGTA	2019	1	24	2	Extraction_Control
Branco08-01	N701	TAAGGCGA	S502	CTCTCTAT	2019	1	18	1	Extraction_Control
branco08-03	N703	AGGCAGAA	S506	ACTGCATA	2019	1	24	1	Extraction_Control
Branco08-03	N706	TAGGCATG	S506	ACTGCATA	2019	2	1	1	Extraction_Control
Branco08-03-dup	N712	GTAGAGGA	S505	GTAAGGAG	2019	2	1	2	Extraction_Control
branco08-03-dup	N709	GCTACGCT	S507	AAGGAGTA	2019	1	24	2	Extraction_Control
Branco-08-11-dup	N707	CTCTCTAC	S503	TATCCTCT	2019	1	18	2	Extraction_Control
Branco11-01	N701	TAAGGCGA	S504	AGAGTAGA	2019	1	18	1	Extraction_Control
branco11-01-dup	N707	CTCTCTAC	S505	GTAAGGAG	2019	1	18	2	Extraction_Control
branco24-02	N705	GGACTCCT	S507	AAGGAGTA	2019	2	1	1	Extraction_Control
branco24-02-dup	N711	AAGAGGCA	S508	CTAAGCCT	2019	2	1	2	Extraction_Control
Branco26-01	N701	TAAGGCGA	S507	AAGGAGTA	2019	1	18	1	Extraction_Control
branco26-01-dup	N707	CTCTCTAC	S508	CTAAGCCT	2019	1	18	2	Extraction_Control
Branco29-02	N704	TCCTGAGC	S504	AGAGTAGA	2019	1	29	1	Extraction_Control
branco29-02-dup	N710	CGAGGCTG	S505	GTAAGGAG	2019	1	29	2	Extraction_Control
CN	N707	CTCTCTAC	S517	GCGTAAGA	2019	5	30	1	PCR_Control
CN-dup	N712	GTAGAGGA	S508	CTAAGCCT	2019	5	30	2	PCR_Control
CN-PCR1	N701	TAAGGCGA	S508	CTAAGCCT	2019	1	18	1	PCR_Control
CN-PCR1-dup	N708	CAGAGAGG	S517	GCGTAAGA	2019	1	18	2	PCR_Control
CN-PCR2	N703	AGGCAGAA	S507	AAGGAGTA	2019	1	24	1	PCR_Control
CN-PCR2-dup	N709	GCTACGCT	S508	CTAAGCCT	2019	1	24	2	PCR_Control
CN-PCR3	N705	GGACTCCT	S502	CTCTCTAT	2019	1	29	1	PCR_Control
CN-PCR3-dup	N711	AAGAGGCA	S503	TATCCTCT	2019	1	29	2	PCR_Control
CN-PCR4	N706	TAGGCATG	S507	AAGGAGTA	2019	2	1	1	PCR_Control
CN-PCR4-dup	N712	GTAGAGGA	S506	ACTGCATA	2019	2	1	2	PCR_Control
MOCK	N706	TAGGCATG	S508	CTAAGCCT	2019	5	30	1	Positive_Control
MOCK-dup	N712	GTAGAGGA	S507	AAGGAGTA	2019	5	30	2	Positive_Control

Anexo C - Legenda da taxonomia obtida para todas as amostras, ao nível de espécie


d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Micrococcales;f__Micrococcaceae;g__Kocuria;s__Kocuria_sp.
d__Bacteria;p__Campilobacterota;c__Campylobacteria;o__Campylobacterales;f__Sulfurovaceae;g__Sulfurovum;s__uncultured_bacterium
d__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Bacillaceae;g__Bacillus;__
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__uncultured;__
d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Lactococcus;s__Lactococcus_lactis
d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Carnobacteriaceae;g__Trichococcus;s__uncultured_bacterium
d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Corynebacteriales;f__Nocardiaceae;g__Nocardia;s__Nocardia_beijingensis
d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacteraceae;__;__
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;__
d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Flammeovirgaceae;g__Rapidithrix;s__Rapidithrix_thailandica
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Gammaproteobacteria_Incertae_Sedis;f__Unknown_Family;g__Unknown_Family;s__uncultured_bacterium
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas;s__Pseudomonas_putida
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Moraxella;s__uncultured_bacterium
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Oxalobacteraceae;g__Herbaspirillum;s__Herbaspirillum_huttiense
d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Rhizobiaceae;g__Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium;__
d__Bacteria;p__Acidobacteriota;c__Vicinamibacteria;o__Vicinamibacterales;f__Vicinamibacteraceae;g__Vicinamibacteraceae;s__uncultured_Acidobacteria
d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus;s__Enterococcus_sp.
d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Corynebacteriales;f__Corynebacteriaceae;g__Corynebacterium;__
d__Bacteria;p__Planctomycetota;c__Planctomycetes;__;__;__;__
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Oceanospirillales;f__Halomonadaceae;g__Halomonas;s__Halomonas_sp.
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacterales;f__Erwiniaceae;g__Pantoea;s__Pantoea_agglomerans
d__Bacteria;p__Actinobacteriota;c__Rubrobacteria;o__Rubrobacterales;f__Rubrobacteriaceae;g__Rubrobacter;s__uncultured_actinobacterium
d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Micrococcales;__;__;__
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Acinetobacter;s__Acinetobacter_calcoaceticus
d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;f__Weeksellaceae;g__Elizabethkingia;s__Elizabethkingia_anophelis
d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Micropepsales;f__Micropepsaceae;g__uncultured;s__uncultured_bacterium
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Rhodocyclaceae;g__Azospira;s__uncultured_bacterium
d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Corynebacteriales;__;__;__
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Diplorickettsiales;f__Diplorickettsiaceae;g__Aquicella;s__uncultured_Aquicella
d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Corynebacteriales;f__Corynebacteriaceae;g__Corynebacterium;s__uncultured_bacterium
d__Bacteria;p__Firmicutes;c__Bacilli;o__Exiguobacterales;f__Exiguobacteraceae;g__Exiguobacterium;s__Exiguobacterium_mexicanum
d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Chitinophagales;f__Saprospiraceae;g__Candidatus_Aquirestis;s__uncultured_bacterium
d__Bacteria;p__Verrucomicrobiota;c__Verrucomicrobiae;o__Pedosphaerales;f__Pedosphaeraceae;g__Pedosphaeraceae;s__uncultured_bacterium
d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Bifidobacteriales;f__Bifidobacteriaceae;g__Bifidobacterium;s__Bifidobacterium_animalis
d__Bacteria;p__Firmicutes;c__Clostridia;o__Lachnospirales;f__Lachnospiraceae;g__Roseburia;s__uncultured_bacterium
d__Bacteria;p__Firmicutes;c__Clostridia;__;__;__;__
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__MBMPE27;f__MBMPE27;g__MBMPE27;s__uncultured_organism
d__Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;__;__;__
d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Frankiales;f__Geodermatophilaceae;g__Geodermatophilus;s__Geodermatophilus_telluris
d__Bacteria;p__Actinobacteriota;c__Actinobacteria;o__Micrococcales;f__Micrococcaceae;g__Arthrobacter;__
d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;__;__
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__Comamonas;s__Comamonas_aquatica
d__Bacteria;p__Planctomycetota;__;__;__;__;__
d__Bacteria;p__Bacteroidota;c__Bacteroidia;__;__;__;__
d__Bacteria;p__Desulfobacterota;c__Desulfobacteria;o__Desulfatiglandales;f__Desulfatiglandaceae;g__Desulfatiglans;s__uncultured_bacterium
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__uncultured;s__Acidovorax_sp.
d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Rhodobacteraceae;g__Rhodobacter;s__uncultured_bacterium
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Chloroplast;f__Chloroplast;g__Chloroplast;s__uncultured_marine
d__Bacteria;p__Armatimonadota;c__Fimbriimonadia;o__Fimbriimonadales;f__Fimbriimonadales;g__Fimbriimonadales;s__uncultured_anaerobic
d__Eukaryota;p__Parabasalia;__;__;__;__;__
d__Eukaryota;__;__;__;__;__;__
d__Bacteria;p__Chloroflexi;c__Anaerolineae;o__SBR1031;f__SBR1031;g__SBR1031;s__uncultured_bacterium
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Chloroplast;f__Chloroplast;g__Chloroplast;s__Solanum_melongena
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas;s__Pseudomonas_oleovorans
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas;s__Pseudomonas_azotoformans
d__Bacteria;p__Fusobacteriota;c__Fusobacteriia;o__Fusobacteriales;f__Fusobacteriaceae;g__Fusobacterium;s__uncultured_organism
d__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Planococcaceae;g__Lysinibacillus;s__Lysinibacillus_sphaericus
d__Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Bacillaceae;__;__

Anexo D –Taxa contaminantes identificados em controlos negativos deste estudo e em múltiplos estudos da literatura.

Género	Controlo Extração	Controlos PCR	Amostras	Referências
Acinectobacter	Presente	Ausente	Presente	[30] [31,32]
Agromyces	Presente	Presente	Ausente	*
Allorhizobium-Neorhizobium-Nararhizobium-Rhizobium	Presente	Presente	Presente	[33]
Arthobacter	Presente	Presente	Presente	[31,34]
Azospira	Presente	Ausente	Presente	[31]
Bacillus	Presente	Presente	Presente	[26,31,34]
Bifidobacterium	Presente	Ausente	Ausente	[28]
Bosea	Presente	Presente	Presente	[31]
Chloroplast	Presente	Ausente	Ausente	[33]
Comamonas	Presente	Ausente	Presente	[26,31,34]
Corynebacterium	Presente	Presente	Presente	[26,31,34]
Elizabethkingia	Presente	Ausente	Presente	[33]
Enterococcus	Presente	Presente	Presente	[26,32,34]
Escherichia-Shigella	Presente	Ausente	Ausente	[28]
Fimbriimonadales	Presente	Ausente	Ausente	*
Halomonas	Ausente	Presente	Ausente	[26]
Kocuria	Ausente	Presente	Presente	[31]
Lysinibacillus	Presente	Presente	Presente	*
Novosphingobium	Presente	Ausente	Ausente	[26,31]
Pasteurella	Presente	Presente	Presente	[26]
Paracoccus	Presente	Ausente	Ausente	[31]
Pseudomonas	Presente	Ausente	Presente	[31,34]
Ralstonia	Presente	Ausente	Presente	[26,31]
Rapidithrix	Presente	Ausente	Ausente	*
Rhodobacter	Presente	Ausente	Ausente	*
S0134_terrestrial_group	Presente	Presente	Ausente	*
Sphingomonas	Presente	Presente	Presente	[26,31]
Staphylococcus	Presente	Presente	Presente	[26,33]
Stenotrophomonas	Presente	Ausente	Presente	[26,30–32]
Streptococcus	Ausente	Presente	Presente	[26,31,32,34]
Streptomyces	Presente	Ausente	Ausente	*
Variovax	Presente	Presente	Presente	[31]
Vermiphilaceae	Presente	Ausente	Ausente	*
Vicinamibacteraceae	Presente	Ausente	Ausente	*

**Não foram encontradas referências na bibliografia consultada relativa a estes taxas, no entanto tal não significa que não sejam reportados como possíveis contaminantes em outros estudos metagenómicos.*